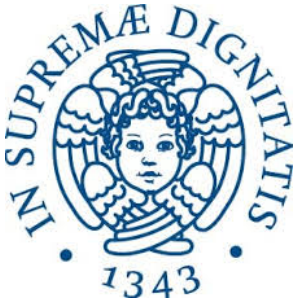


UNIVERSITÀ DI PISA

Dipartimento di Informatica
Laurea Magistrale in Informatica per l'Economia e per l'Azienda
(Business Informatics)

TESI DI LAUREA

PROGETTAZIONE E REALIZZAZIONE DI UN DATAWAREHOUSE PER IL CONTROLLO DELLE VENDITE E L'ANALISI DEL CLIENTE NEL COMMERCIO ELETTRONICO

Candidato
Alessandro Di Paola

Relatori
Prof. Salvatore Ruggieri
Dott. Alfonso Baccari

Anno Accademico 2013/14

Riassunto

Si presenta un progetto di Business Intelligence per il disegno e la realizzazione di un Data Warehouse di supporto ai processi aziendali di controllo delle vendite e analisi del cliente nel commercio elettronico. L'attenzione è rivolta verso l'attività di integrazione tra le sorgenti dati provenienti da diversi sistemi quali il commercio elettronico e il Data Warehouse esistente. Dopo aver presentato l'ambito in cui si è svolto il lavoro alla base di questa tesi, ed introdotto il concetto di Web Analytics, sono esposte le fasi di analisi dei requisiti, di progettazione e realizzazione dei Data Mart e del Data Warehouse finale. Sono così descritte le fasi di estrazione, trasformazione e caricamento dei dati nei sistemi analitici progettati, per finire con la descrizione della reportistica creata. Di ogni fase saranno presentate sia le problematiche di ordine generale, sia le soluzioni ai problemi riscontrati durante l'esperienza diretta con la realtà aziendale.

Indice

Capitolo 1

Introduzione

1.1 Presentazione del problema

Il commercio elettronico nel 2014 ha raggiunto i 1.500 miliardi di dollari, crescendo del 20% rispetto all'anno precedente. La crescita di tale fenomeno è dovuta in gran parte alla diffusione di smartphone e tablet rendendo il *mobile* fattore chiave di successo per il commercio elettronico. Le aziende operanti nel mercato online che hanno colto le opportunità portate dal nuovo tipo di accesso sono state in grado di sviluppare e ridefinire la propria strategia di business, mentre le aziende che non hanno affrontato questo cambiamento hanno avuto un calo del *conversion rate* sul proprio sito.

Espandere il proprio canale di vendita nel Web ha permesso ai negozi fisici di ampliare la propria clientela e ai produttori di beni e servizi di eliminare nella catena di vendita le aziende di distribuzione. Questo tipo di integrazione viene sfruttato nel campo della moda. La crescita di questo settore infatti, è dovuta allo sviluppo del commercio elettronico che ha affiancato un sistema di vendita al dettaglio già consolidato. Le innovazioni tecnologiche permettono di trasformare il negozio da un luogo di stoccaggio e vendita ad uno spazio integrato digitale, progettato per offrire la massima funzionalità, coinvolgere e divertire [Casaleggio 2014]. La promozione online del brand e lo sviluppo della presenza sul Web, secondo l'indagine qui riportata, è ancora un'attività difficoltosa per la maggior parte delle aziende italiane. Come evidenzia [Kaushik 14], gli analisti e gli esperti di marketing hanno una visione molto limitata dei dati relativi al Web e al loro utilizzo. In questo contesto si inserisce la Web Analytics, che ha rivoluzionato il modo di utilizzare i dati del Web per aiutare a conseguire gli obiettivi strategici di business.

Applicando i principi della Business Intelligence è possibile integrare le informazioni provenienti dal sistema gestionale aziendale con le informazio-

ni ricavate dalla raccolta, conservazione ed elaborazione dei dati Web. Il risultato di questo processo è un Sistema di Supporto alle decisioni in grado di creare rapporti informativi per il controllo direzionale, ovvero di rispondere alle domande sulle tendenze in atto, sui risultati e sulle modalità di svolgimento dell'attività online.

Il lavoro da me svolto all'interno di *ICONSULTING S.p.A.* si è focalizzato sullo sviluppo di un sistema di supporto decisionale nell'ambito del commercio elettronico. L'obiettivo è stato quello di sfruttare i benefici derivanti dallo strumento che si è rivelato in grado sia di analizzare gli scostamenti economici dell'attività di vendita online, sia di analizzare i comportamenti dei visitatori sul sito web.

Il committente del progetto è rappresentato da una multinazionale italiana operante nel mercato degli occhiali, ed in particolare dal settore che all'interno dei processi aziendali è chiamato a prendere decisioni a cadenza periodica relative alle attività di *web analysis* per pianificare interventi sul sito e ottimizzare il ritorno degli investimenti. L'obiettivo del progetto, descritto in questo lavoro di tesi, è la realizzazione di un data warehouse che raccolga i dati generati dall'attività online dei visitatori e dalle vendite per la realizzazione di una reportistica finalizzata al monitoraggio delle performance del sito web attraverso indicatori chiave.

1.2 Rassegna della letteratura

Di seguito sono presentati i testi consultati per la stesura della tesi:

- Per la parte introduttiva riguardante la Web Analytics: sono stati fondamentali i testi di [Ruggieri 13], [Kimball 00], [WAA 07] e [Kaushik 14];
- Per la parte relativa alla progettazione del DataWarehouse e le soluzioni proposte ai ricorrenti problemi di natura logica e concettuale si è fatto riferimento alle soluzioni presenti in [Albano 13] e [Kimball 96];
- Per la parte implementativa sono state utilizzate a seconda dei problemi affrontati, le dispense presenti in [Icowiki 14] e il testo di [Kimball 96];
- Per la documentazione degli strumenti utilizzati si è fatto ampiamente riferimento a [Sap 14], [Netezza 09] e [Garcia 12].

1.3 Contenuto della tesi

Il lavoro che ha ispirato questa tesi è stato svolto presso l'azienda bolognese *ICONSULTING S.p.A.* uno dei maggiori System Integrator indipendenti italiani specializzata unicamente in progetti di Data Warehouse, Business Intelligence e Corporate Performance Management. L'obiettivo del progetto, realizzato per un'azienda del settore ottico, di cui quest'opera descrive le fasi di disegno e di sviluppo, è la creazione di un Data Warehouse al fine di creare un modello integrato a supporto delle decisioni aziendali. Di seguito viene presentata l'organizzazione della tesi indicando, per ogni capitolo, il proprio contenuto.

- Nel **Capitolo 2** si presenta la disciplina alla base di questo progetto di tesi: la Web Analytics. Si descrivono, inoltre, gli elementi alla base dell'analisi dei dati Web; il capitolo termina illustrando come la Web Analytics possa integrarsi in un sistema di supporto alle decisioni.

- Il **Capitolo 3** si apre con la descrizione dell'azienda committente, continua con l'illustrazione delle esigenze che hanno portato alla realizzazione del progetto in questione ed espone i processi aziendali coinvolti. Infine sono proposti i requisiti di analisi raccolti e la verifica dei dati a disposizione.

- Nel **Capitolo 4** si presenta la prima fase di progettazione del Data Warehouse per il monitoraggio dei processi descritti nel precedente capitolo. Una volta definite le specifiche si descrive la progettazione concettuale iniziale, poi si passa alla progettazione concettuale basata sui dati operazionali fino ad arrivare alla progettazione concettuale finale.

- Nel **Capitolo 5** viene affrontata la fase conclusiva della progettazione attraverso la presentazione delle sorgenti informative e la rappresentazione logica dello schema relazionale del Data Warehouse.

- Nel **Capitolo 6** vengono descritte le tecnologie e il software sviluppato utilizzato nella realizzazione delle fasi di estrazione e caricamento dei dati e fino allo sviluppo dei report. Per ogni tecnologia è riportata un'introduzione sulle principali caratteristiche tecniche e sulla sua architettura.

- Nel **Capitolo 7** viene descritto lo sviluppo delle procedure di estrazione, trasformazione e caricamento dei dati. Si delineano i punti salienti della implementazione, ponendo l'attenzione sulla descrizione dei problemi riscontrati nell'affrontare la realizzazione di tali procedure e le relative soluzioni adottate.

- Nel **Capitolo 8** si presenta la descrizione della reportistica prodotta suddivisa per destinazione d'uso.

La tesi termina con la descrizione degli obiettivi raggiunti, le conclusioni derivate dall'esito del lavoro svolto ed alcuni cenni sulle proposte per ulteriori sviluppi futuri.

Capitolo 2

Web Analytics

Il presente capitolo inizia con una breve introduzione alla Web Analytics, prosegue approfondendo il tema e ponendo l'accento su come questa disciplina possa essere di supporto al conseguimento degli obiettivi di business. Nel corso del capitolo verranno descritti gli elementi base del flusso dei dati Web, ovvero le metriche e gli indicatori chiave di performance, e per ultimo verrà presentata una breve esposizione su come gli strumenti di misurazione ed analisi della Web Analytics possano creare valore per l'azienda.

2.1 Introduzione

Una chiara definizione di Web Analytics è presente in [WAA 07]

Definition 1 (Web Analytics) *La Web Analytics è la misurazione, la collezione, l'analisi e il reporting di dati Web, allo scopo di capire ed ottimizzare l'utilizzo del sito.*

Mediante metriche quantitative e qualitative, la Web Analytics mira a profilare l'utente, a modellare il sito in base alle sue esigenze, con il fine ultimo di migliorare la sua esperienza (Web experience). Attraverso la Web Analytics le imprese che operano sul Web possono analizzare un'immensa fonte di dati e monitorare il comportamento degli utenti per supportare il processo decisionale e distribuire prodotti/servizi migliori e personalizzati.

2.2 La Web Analytics come strategia di Business

Attualmente molte aziende riversano ingenti risorse sui nuovi media digitali e sono sempre più attente ad analizzare e comprendere quale sia il ritorno dell'investimento. Il grande vantaggio del Web Marketing è quello di permettere una misurazione immediata e di poter avere un quadro preciso di come si stia sviluppando un progetto online. La Web Analytics è il passaggio fondamentale per avere il controllo della situazione. Non è un caso che

dietro ad aziende di successo online vi siano persone e organizzazioni che adottano la Web Analytics come strategia di impresa [Kaushik 14]. La Web Analytics come strategia di business, è intesa come il processo continuo di allineamento dell'impresa all'evoluzione del mercato online.

Il processo strategico avviene:

- allocando risorse economiche, competenze nell'analisi dei dati Web e investendo sulle tecnologie e sugli strumenti dedicati alla Web Analytics;
- sfruttando le opportunità e riducendo l'impatto delle minacce provenienti dall'esterno, attraverso soluzioni di *Competitive Intelligence*, ovvero analisi dei dati Web dei concorrenti.

Allo stato attuale del settore, la maggior parte delle imprese considera la Web Analytics come l'arte di raccogliere e analizzare i dati dei clic forniti da strumenti proprietari. Anche se un buon punto di partenza, ci si rende conto di quante poche informazioni utili si riescano a ottenere rispetto alla mole di dati in possesso. Secondo quanto riportato in [Kaushik 14], per assimilare la Web Analytics come strategia di business è necessaria una ridefinizione del concetto di analisi dei dati Web, che consiste:

- nell'analisi dei dati quantitativi e qualitativi tratti dal sito Web proprio e della concorrenza;
- nel continuo miglioramento dell'esperienza online per i clienti attuali e potenziali;
- nel confronto tra i risultati rilevati e quelli attesi.

In questo modo, l'esperto di marketing online è obbligato ad ampliare il suo bagaglio di conoscenze e di strumenti in possesso, rivedendo la Web Analytics come processo strategico a lungo termine suddiviso nelle seguenti fasi: acquisizione dei dati Web (Clickstreaming), analisi dei risultati, sperimentazione e test, interviste ai clienti e Competitive Intelligence.

Nella restante parte del capitolo verranno approfonditi i primi due elementi, fondamentali per lo sviluppo del progetto presentato in questa tesi.

2.3 Il Clickstream

Definition 2 (Clickstreaming) *Il* Clickstreaming *è la raccolta, la conservazione, l'elaborazione e l'analisi dei dati a livello dei clic.*

I dati a livello di clic aiutano a valutare l'efficacia di pagine web, delle campagne pubblicitarie e ad analizzare tutti i tipi di comportamento del sito: visite, visitatori, tempo trascorso sul sito, pagine visualizzate, frequenza di

rimbalzo ecc [Kaushik 14]. Nel seguente paragrafo sono descritti le metriche e gli indicatori chiave di prestazioni, i quali costituiscono gli elementi base dell'analisi dei dati Web.

2.3.1 Le metriche e gli indicatori chiave di prestazioni

Una metrica è una misurazione statistica e quantitativa che descrive gli eventi o le tendenze in atto su un sito Web. Un indicatore chiave delle prestazioni è una metrica che aiuta a confrontare le prestazioni in analisi con gli obiettivi aziendali, specifici per ogni azienda.

Di seguito sono presentate le metriche principali, approfondendo il concetto di Visita a Visitatore Unico che costituiscono la base di ogni calcolo di metriche Web.

Visita e Visitatore Unico Tecnicamente la visita è chiamata *sessione*. La sessione è una raccolta di richieste prodotte dall'utente al sito Web.

Quando l'utente richiede la prima pagina o il primo elemento del sito, lo strumento di analisi avvia una sessione specifica per tale persona e tale browser. Ogni ulteriore richiesta di questo utente viene attribuita a questa sessione univoca. Quando la persona lascia il sito, il suo codice univoco di sessione viene utilizzato per raggruppare le pagine visualizzate in un'unica visita.

Quando si esegue un report per un determinato periodo, tramite gli strumenti di Web Analytics, le visite totali rappresentano *il numero di tutte le sessioni registrate nell'arco di tale periodo.*

La metrica Visitatore Unico è un'approssimazione del numero di persone che visitano il sito. Quando l'utente richiede la prima pagina o il primo elemento del sito, lo strumento di analisi crea un file, chiamato *cookie*, con all'interno un codice univoco. Il file rimane nel browser anche dopo che l'utente abbandona il sito. Ogni volta che qualcuno visita il sito Web, il cookie viene utilizzato per riconoscere se un utente è nuovo o è ritornato ad iniziare un'altra Visita.

Quando si esegue un report per un determinato periodo temporale con lo strumento di Web Analytics, la metrica Visitatore Unico rappresenta *il numero di tutti i cookie persistenti univoci rilevati nell'arco di tale periodo.*

Time on Page e Time on Site Sono rispettivamente il tempo che i Visitatori trascorrono in una singola pagina e il tempo totale trascorso sul sito nel corso di una sessione.

Bounce Rate È definito come la percentuale delle sessioni sul sito web in cui è stata visualizzata una sola pagina. Quando si esegue un report,

è possibile misurare il Bounce Rate a livello dell'intero sito e delle singole pagine web. Nel primo caso mostra il grado di fallimento del sito. Nel secondo caso, aiuta ad identificare le pagine (es. le home page) che non funzionano e provocano troppi rimbalzi.

Exit Rate È definito come il numero di persone che sono uscite dal sito Web da una determinata pagina. Il tasso di uscita mostra la percentuale di coloro che sono entrati in qualche modo nel sito, ma sono usciti da una determinata pagina;

Conversion Rate Il Conversion Rate, espresso come percentuale, è il rapporto tra *risultati* e Visitatori Unici (o Visite). I risultati normalmente sono rappresentati dall'invio di un ordine al sito Web di commercio elettronico. La scelta del denominatore, invece, dipende da quale modello di analisi è adatto per la propria attività. La metrica Visita si usa per quei siti in cui lo stesso Visitatore potrebbe svolgere più acquisti in un breve arco di tempo. Se si utilizza la metrica Visitatore Unico, si suppone che qualcuno visiti il sito più volte prima di effettuare un acquisto. Utilizzando quest'ultima metrica, il calcolo del tasso di conversione rispecchia l'intero processo decisionale di acquisto del cliente.

Qualità di una metrica

Le metriche, indipendentemente dalla loro efficacia, devono perdurare nel tempo e ai cambiamenti delle attività. [Kaushik 14] propone quattro qualità da attribuire ad una bouna metrica.

- **Semplice** Per favorire un'azione aziendale, le prestazioni e le decisioni devono essere comprensibili e prive di complicazioni.
- **Rilevante** La metrica deve essere rilevante nel contesto in cui viene usata. Ovvero le metriche identificate devono essere rilevanti per misurare obiettivi di successo specifici per l'azienda e per il sito Web in possesso.
- **Tempestiva** Le metriche migliori arrivano tempestivamente, in modo che chi è incaricato di prendere decisioni in azienda, possa prendere decisioni tempestive anche se a discapito della complessità e della perfezione.
- **Utile istantaneamente** Una metrica è utile istantaneamente quando si comprende il suo significato e se fornisce informazioni utili a primo impatto.

2.4 Web Analytics e Business Intelligence

Di seguito sono riportate le varie fasi che fanno della Web Analytics un processo decisionale di Business Intelligence, trasformando i dati e le informazioni raccolte in "conoscenza". Come ogni sistema di Business Intelligence, anche la Web Analytics ha un obiettivo preciso, che dipende dalla visione aziendale e dagli obiettivi posti dalla gestione strategica di un'azienda. A tal proposito le fasi proposte sono:

1. **Strategia.** Discussa nel paragrafo 2.2;

2. **Selezione.** Associare gli obiettivi strategici a breve termine, definiti nella strategia di business, con i parametri, le metriche e i KPI.

   L'attività di misurazione inizia identificando i fattori critici di successo: bisogna determinare un'insieme di possibili metriche che riflettano i fattori individuati (ad esempio capacità di acquisizione del cliente, margini di profitto, customer satisfaction ecc.), separare le informazioni interessanti dalle informazioni indispensabili e individuare quei KPI in grado di sintetizzare l'andamento dell'azienda. Alla fine di questa fase saranno messe in luce le metriche minime indispensabili che danno agli interessati del processo decisionale un senso di concentrazione e direzione.

3. **Analisi.** Monitoraggio e misurazione degli indici di prestazioni selezionati al passo 2 attraverso la creazione di una o più dashboard dando alla direzione aziendale le informazioni di cui hanno bisogno per prendere decisioni corrette. Ogni *dashboard* può essere finalizzata per uno specifica analisi, ad esempio: analisi per l'ottimizzazione dei motori di ricerca, analisi del traffico diretto, analisi delle campagne di posta elettronica ecc.

4. **Ottimizzazione.** Le analisi costruite al passo 3 permettono di intraprendere un processo di controllo e miglioramento continuo quali: miglioramento delle ricerche interne al sito, ottimizzazione delle indicizzazioni per i motori di ricerca, ristrutturazione di pagine web che presentano un Bounce Rate elevato ecc.

Alle attività sopra esposte se ne aggiunge una trasversale, ovvero una *fase di integrazione* dei dati Web con i dati globali che rappresentano l'azienda nel suo complesso e che provengono da tutto il contesto generale. L'integrazione dei dati è spinta dall'esigenza di comprendere l'impatto e il valore economico del sito Web in possesso.

Capitolo 3

Caso di studio

Il presente capitolo è incentrato sulla fase di raccolta dei requisiti. Viene presentata l'azienda per la quale è stata realizzato il progetto, proseguendo con l'analisi dei processi aziendali per individuarne i più importanti, le loro priorità e quale informazione si ritiene utile produrre con l'analisi dei dati. Si descrivono quindi, i requisiti di analisi raccolti e i problemi riscontrati in fase di integrazione tra le varie sorgenti informative. Infine viene brevemente descritto l'incontro con il reparto IT per verificare la fattibilità dei requisiti raccolti, sia in termini di disponibilità informativa che a livello operazionale.

3.1 L'azienda committente

In questo paragrafo viene descritta la natura e la missione dell'impresa committente e successivamente gli attori e le logiche, caratteristiche dell'azienda in oggetto, focalizzando l'attenzione sull'attività del commercio elettronico.

3.1.1 Analisi della natura e dei fini dell'azienda

L'azienda committente opera nella produzione e distribuzione di occhiali, attualmente è il più grande produttore di lenti e montature. La missione dell'azienda è offrire, ad una clientela mondiale, occhiali da sole e da vista di elevata qualità tecnica e stilistica al fine di migliorare il benessere e la soddisfazione dei propri clienti, creando nel contempo valore per i dipendenti e la comunità in cui opera.

3.1.2 Organizzazione distributiva

La distribuzione diretta permette all'impresa di proporre i suoi prodotti nei principali mercati, e di identificare in modo univoco i gusti e le tendenze dei consumatori finali. A tal fine, l'impresa si avvale principalmente di due canali: *distribuzione retail*, ovvero la vendita al dettaglio degli opera-

tori commerciali ai consumatori finali e *distribuzione wholesales*, ovvero la vendita all'ingrosso.

Il sistema distributivo integrato si appoggia a un apparato centrale di pianificazione della produzione. La rete che raccorda i centri logistici e di vendita con gli impianti produttivi consente di controllare giornalmente l'andamento delle vendite nel mondo e i livelli di scorte, programmando le risorse produttive e riallocando prontamente le scorte di magazzino di base alla specifica domanda dei singoli mercati.

Grazie all'efficienza logistica e alla notevole flessibilità in fase di produzione e assemblaggio, è stato possibile intraprendere, nell'ambito del commercio elettronico, una strategia di produzione basata sulla personalizzazione di massa, creando per il cliente, una piattaforma online di personalizzazione del prodotto. I siti Web di commercio elettronico sono complementari alle attività retail legate al marchio e alla distribuzione internazionale, essi permettono ai clienti di acquistare prodotti nel modo più efficiente possibile aumentando nel contempo la riconoscibilità dei marchi, migliorando il servizio al cliente e veicolandone adeguatamente i valori e l'essenza.

3.2 Analisi dei processi aziendali

Sono stati effettuati diversi incontri, con i delegati della direzione commerciale prima e con i responsabili del reparto IT poi, per comprendere quali siano i processi chiave che necessitano di uno strumento analitico per il supporto alle decisioni e quali sorgenti informative sono a disposizione per soddisfare queste esigenze. Inoltre viene identificata il tipo di informazione che attualmente manca a livello decisionale e che si ritiene utile produrre con l'analisi dei dati. Per ogni processo identificato, vengono raccolti i requisiti di analisi dei dati richiesti. Infine, vengono delineate le misure specifiche di ciascun requisito.

3.2.1 Processo di vendita

Il processo di vendita online rientra nella fase di acquisto del prodotto da parte del cliente.

Affinché l'ordine venga formalizzato, il cliente deve specificare il metodo di pagamento, l'utilizzo o meno di codice promozionale e i dati di spedizione. L'ordine di vendita è evaso solo se approvato e solo se il pagamento è andato a buon fine. I tempi di consegna sono generalmente brevi, si considerano dai 3 ai 4 giorni lavorativi. Nel caso in cui il prodotto in spedizione non fosse disponibile, il cliente può:

- accettare il ritardo;
- cancellare totalmente l'ordine, nel caso in cui comprenda un solo prodotto, ed effettuare contestualmente l'acquisto di un nuovo prodotto;

- cancellare totalmente l'ordine, nel caso in cui comprenda un solo prodotto, e avere un riaccredito completo;
- cancellare parzialmente l'ordine, nel caso comprenda più prodotti, per il solo prodotto che dovesse risultare non disponibile, e procedere con il resto degli acquisti che verranno consegnati nei tempi previsti.

Nel caso in cui il prodotto fosse disponibile, la procedura non prevede la possibilità di cancellare l'ordine da parte del cliente, ma una volta consegnato, il cliente può effettuare un reso entro 45 giorni lavorativi dalla data di consegna. Per restituire uno o più prodotti è necessario un'autorizzazione al reso e dal momento della ricezione di conferma dell'autorizzazione, il cliente ha 30 giorni per effettuare la spedizione del reso.

Le spese di spedizione sono associate ad un intero ordine. La fatturazione avviene nel momento in cui i prodotti sono spediti presso la sede del cliente, quindi, ogni spedizione è accompagnata da una fattura.

Durante la raccolta dei requisiti sono emerse alcune considerazioni; i delegati della direzione valutano il successo di un prodotto tramite lo scostamento del *ricavo totale delle vendite giornaliere dei prodotti*, al netto dello sconto ad essi associati, per prodotto, per giorno e per cliente rispetto all'anno precedente o ad un valore target pianificato a preventivo.

Nel seguito useremo la seguente terminologia:

- I prodotti hanno un prezzo unitario. Quando un prodotto è venduto in una certa quantità, interessano le seguenti grandezze:

  $prezzo = prezzo\ unitario \times quantità\ venduta$

- Il ricavo totale del venduto, della vendita di un prodotto in una certa quantità, ad un prezzo ridotto per effetto di uno sconto, è definito come:

  $ricavo = prezzo$ - $sconto$

Di ogni vendita giornaliera di un prodotto si conosce il totale delle seguenti grandezze: il prezzo, la quantità venduta, lo sconto e il numero di ordini.

Raccolta dei requisiti del Processo di vendita

N	Requisito
1	Valore dei ricavi per prodotto, cliente e data
2	Valore medio degli ordini spediti per prodotto, cliente e data
3	Valore medio di unità venduta per prodotto, cliente e data

Tabella 3.1: Requisiti di analisi per il processo Vendita

Per semplificare la scrittura dei requisiti di analisi, oltre a restare valide le considerazioni fatte in tabella, si è supposto che:

- Con l'analisi per prodotto si intende la neccessità di analizzare per codice per prodotto, brand, categoria e collezione;
- Con l'analisi per data si intende la possibilità di analizzare per periodo fiscale o solare.

Le misure emerse dall'analisi sono: Venduto, Ordinato e Unità.

3.2.2 Processo di navigazione dell'utente

Il processo o lo scenario che si presenta all'utente nel momento in cui accede al sito di commercio elettronico può essere definito come la raccolta di richieste prodotte durante una Visita.

Sono stati selezionati tre scenari che riassumono i vari modelli comportamentali dei visitatori:

1. **Conversione.** Definito come il processo di trasformazione dei visitatori in clienti paganti.
2. **Abbandono.** Il visitatore decide di acquistare uno o più prodotti sul sito, ma durante il processo di acquisto rinuncia.
3. **Rimbalzo.** Il visitatore che accede ad una qualsiasi pagina web del sito di commercio elettronico esce senza visitare altre pagine.

Tralasciando il terzo scenario che si conclude immediatamente, andiamo ad analizzare i primi due, che presentano una fase preliminare comune. In entrambi, si suppone che il Visitatore accede alla pagina Web di interesse, ovvero alle informazioni di carattere generale sulle caratteristiche del prodotto e seleziona quello adatto alle sue necessità. Infine, l'utente aggiunge il prodotto al carrello virtuale. A questo punto i due scenari proseguono in direzioni diverse:

1. **Conversione.** Il visitatore sceglie se accedere come cliente registrato o proseguire l'acquisto senza registrarsi; compila il dettaglio indirizzo di spedizione e il dettaglio di indirizzo di fatturazione; sceglie il metodo di pagamento e conferma l'ordine.

2. **Abbandono.** Durante il processo di acquisto l'utente, per qualche motivo rinuncia all'acquisto svuotando il carrello e/o abbandonando il sito.

Durante la raccolta dei requisiti sono emerse alcune considerazioni; i responsabili del reparto di vendita online sono interessati alla popolarità di un prodotto, ovvero al numero di *visitatori unici* che hanno contribuito al totale delle sue vendite giornaliere; per questa ragione interessa conoscere anche il numero di ordini giornalieri con il quale è stato acquistato un prodotto nel sito di commercio elettronico.

Raccolta dei requisiti del Processo di navigazione

N	Requisito
1	Totale delle visite per prodotto, canale di marketing, chiave di ricerca, cliente, dispositivo e data
2	Totale dei visitatori unici per prodotto, canale di marketing, chiave di ricerca, cliente, dispositivo e data
3	Totale degli ordini per prodotto, canale di marketing, chiave di ricerca, cliente, dispositivo e data
4	Numero delle pagine visualizzate per prodotto, canale di marketing, chiave di ricerca, cliente, dispositivo e data
5	Conversion rate per prodotto, canale di marketing, chiave di ricerca, cliente, dispositivo e data
6	Bounce rate per prodotto, canale di marketing, chiave di ricerca, cliente, dispositivo e data

Tabella 3.2: Requisiti di analisi per il processo di navigazione

Per semplificare la scrittura dei requisiti di analisi, oltre alle considerazioni fatte in precedenza, si è supposto che:

- Con l'analisi per prodotto si intende la possibilità di analizzare per: codice prodotto, brand, categoria e collezione;
- Per ogni prodotto esiste una pagina web associata;
- Per canale di marketing si intende il tipo di campagna di marketing che ha generato traffico Web (Email mktg, Brand Display, Paid Search, Social Media ecc.);

- Per dispositivo si intende il tipo di apparecchiatura usata dall'utente per accedere al sito (PC, smartphone e tablet);
- Per chiave di ricerca si intende la parola utilizzata nei motori di ricerca per raggiungere il sito;
- Il denominatore utilizzato per il calcolo del conversion rate è il Visitatore Unico.

Le misure emerse dall'analisi sono: Visita, Visitatore Unico, Ordine.

3.3 Verifica dei dati operazionali

Dopo aver raccolto i requisiti di analisi sono stati organizzati degli incontri con i responsabili IT. In questi colloqui si è discusso riguardo il fabbisogno informativo di ciascun requisito di analisi e, per ognuno di essi, se il fabbisogno informativo poteva essere soddisfatto dai dati operazionali presenti a sistema. Il risultato della verifica è che:

- la raccolta e la conservazione del flusso dei dati Web è gestito da *Omniture Site Catalyst*, fornitore esterno di strumenti per la Web Analytics;
- i dati relativi agli ordini di vendita online sono conservati e gestiti dal sistema *IBM Websphere Commerce Suite* e disponibili nel database Oracle;
- le informazioni sui prodotti commercializzati e i dati sugli ordini fatturati, sono disponibili nel sistema gestionale presente nel database SAP;

Solo i primi due sistemi prenderanno parte alla creazione del Data Warehouse finale. Il sistema gestionale presente nel DB SAP sarà direttamente integrato a livello front-end, ovvero nel sistema di reportistica.

Verificata la fattibilità tecnica dei requisiti di analisi, si procede con la specifica dei requisiti.

Capitolo 4

Specifica dei requisiti di analisi e progettazione iniziale dei data mart

La progettazione di un Data Warehouse è organizzata in fasi. Essa inizia con l'analisi dei requisiti, prosegue con la progettazione concettuale dei Data Mart fino alla definizione del progetto logico che fornisce gli schemi relazionali dei singoli Data Mart che poi saranno integrati dando vita allo schema del Data Warehouse. In questo capitolo si formula una più schematica descrizione dei requisiti fin qui raccolti, con lo scopo di realizzare una documentazione formale necessaria alla fase successiva di progettazione logica del Data Warehouse presentata nel capitolo successivo.

4.1 Metodologia di progettazione

Un *Data Warehouse* è il database analitico che costituisce le fondamenta di un sistema di supporto alle decisioni, ovvero è un sistema in grado di fornire chiare informazioni agli utenti, in modo che essi possano analizzare dettagliatamente una situazione e prendere facilmente le opportune decisioni sulle azioni da intraprendere. In altre parole, un sistema di supporto alle decisioni è un sistema informativo dedicato ad aiutare i responsabili del business in tutta la gestione d'impresa. Questo sistema deve quindi avere un unico obbiettivo: contribuire ad incrementare le performance economiche aziendali, attraverso la fornitura di informazioni strategiche in modo semplice e veloce. Una chiara definizione di Data Warehouse è proposta da [Albano 13]

Definition 3 (Data Warehouse) *Un Data Warehouse è una raccolta di dati storici integrati, non volatile, organizzata per soggetti e finalizzata al recupero di informazioni di supporto ai processi decisionali*

- **Orientata ai soggetti di interesse**. I dati conservati in un Data Warehouse sono organizzati per tema, anzichè per applicazioni, come avviene per i database operazionali, i quali invece hanno lo scopo di ottimizzare l'elaborazione delle transazioni.

- **Integrati**. Un *Data Warehouse* colleziona informazioni provenienti da diverse fonti e le integra tra loro mediante l'esecuzione di un dispendioso e ben pianificato processo di caricamento e trasformazione dei dati che in differenti sorgenti potrebbero essere rappresentati in formati differenti o potrebbero contenere errori sintattici o semantici.

- **Tempificata**. Mentre i database operazionali conservano solo i dati più recenti che descrivono una particolare entità, un Data Warehouse conserva dati storici al fine di poter analizzare i cambiamenti nel tempo o il trend di un particolare evento.

- **Statico**. La modalità di interrogazione di un Data Warehouse è mirata all'estrazione di dati, non alla modifica dei dati conservati. Questo fa si che i dati siano consistenti nel tempo e periodicamente aggiornati con l'aggiunta dei dati più recenti.

- **Supporto alle decisioni**. Dato che il fine ultimo è l'estrazione di informazioni significative, il design deve essere specificatamente realizzato al fine di ottenere la risposta alle query ritenute di interesse dal soggetto di interesse.

Spesso il Data Warehouse viene partizionato in sottoinsiemi logici chiamati Data Mart. Una definizione di questo termine è riportato in [Icowiki 14]

Definition 4 (Data Mart) *Un Data Mart è un database analitico progettato per soddisfare le esigenze di una specifica funzione di un business (ad esempio, marketing, vendite e finanze).*

Il Data Mart rappresenta un sottoinsieme dei dati contenuti nel Data Warehouse, segue le stesse regole di progettazione, e similmente al Data Warehouse, può contenere dati aggregati e dati di livello base, a seconda di quelle che sono le esigenze delle specifico gruppo di utilizzatori.

I dati vanno organizzati tenendo presente il modo in cui i dirigenti li utilizzano per i propri scopi:

- I dirigenti sono interessati ad analizzare collezioni di *fatti* che riguardano particolare fenomeni aziendali. Ogni fatto è caratterizzato da un insieme di *misure* che sono attributi numerici che riguardano una prestazione o il comportamento di un fenomeno aziendale.

- I dirigenti sono interessati ad analizzare le misure dei fatti secondo prospettive diverse di analisi, o *dimensioni*, per valutare i risultati del

business nel contesto aziendale al fine di trovare soluzioni ai problemi critici o per cogliere nuove opportunità.

- I dirigenti sono interessati ad analizzare i fatti per valutare l'andamento delle prestazioni aziendali sulla base di una serie di *indicatori*, variabili quantitative calcolate con aggregazioni delle misure, che rappresentano al meglio la strategia e che possono dunque essere interpretati come critici per il successo attuale e futuro dell'impresa.

- I dirigenti sono interessati ad analizzare i fatti a diversi livelli di dettaglio per approfondire l'analisi di situazioni interessanti. È utile rappresentare non solo le dimensioni di analisi, ma anche le *gerarchie dimensionali* che interessano gli attributi delle dimensioni.

Seguendo il metodo riportato in [Abano 13] per la progettazione del Data Warehouse si cerca di procedere considerando sia i requisiti di analisi che la base di dati operazionale a disposizione. La progettazione di un Data Warehouse, come accade per le basi di dati, è un'attività complessa organizzata nelle seguenti fasi:

- **Analisi dei requisiti**. Si raccolgono e si definiscono con i committenti i requisiti delle analisi dei dati, di supporto alle decisioni, che si desiderano eseguire.

- **Progettazione concettuale**. Si definisce un modello concettuale dei dati da analizzare.

- **Progettazione logica**. Si trasforma il modello concettuale in un modello logico usando il modello dei dati di un sistema per la gestione di *Data Warehouse*.

- **Progettazione fisica**. Si definiscono le strutture di memorizzazione (indici e viste materializzate) per agevolare le operazioni di analisi dei dati.

Specifica dei requisiti La fase di analisi dei requisiti si suddivide in due sottofasi principali. La Raccolta dei requisiti, che è stata già esposta nel capitolo 3 durante la presentazione del caso di studio e la Specifica dei requisiti, che produce una descrizione dei requisiti di analisi dei dati che ne evidenzi le caratteristiche salienti da modellare poi con la progettazione concettuale. Ciascun evento, derivante dai processi aziendali, è descritto dalle seguenti tabelle:

- Tabella della specifica dei requisiti di analisi: in cui si formalizzano le analisi che si realizzeranno evidenziando le dimensioni, le misure e le metriche coinvolte.

- Tabella del fatto: in cui si descrive il fatto da modellare nel Data Warehouse.
- Tabella delle dimensioni: in cui si descrivono le dimensioni di analisi del fatto.
- Tabella degli attributi dimensionali: in cui si descrivono gli attributi di ogni dimensione, una tabella per ogni dimensione.
- Tabella delle gerarchie dimensionali: in cui si descrivono le gerarchie delle tabelle dimensionali.
- Tabella delle dimensioni che cambiano: in cui si specifica quale strategia viene adottata per le *slowly changing dimensions*.
- Tabella delle misure del fatto: in cui si descrivono le misure del fatto.

Nelle seguenti sotto sezioni saranno descritte le misure e le dimensioni dedotte dalle specifiche riportate e necessarie per riuscire a soddisfare le esigenze preliminari di analisi del committente. Durante la specifica dei requisiti raccolti sono emerse nuove opportunità di analisi, che saranno discusse ed eventualmente integrate durante la progettazione concettuale finale descritta nel capitolo successivo.

4.2 Ordini di Vendita

4.2.1 Specifica dei requisiti del fatto

Di seguito si evidenziano le dimensioni, gli attributi e le misure utilizzate in ogni requisito di analisi inizialmente raccolto relativo agli ordini di vendita.

N	Requisito	Dimensioni	Misure	Aggregazioni
1	Valore dei ricavi per prodotto, cliente e data	Prodotto, Data, Testata Ordine, Riga Ordine, Cliente	Prezzo	Somma
2	Valore medio degli ordini spediti per prodotto, cliente e data	Prodotto, Data, Testata Ordine, Riga Ordine, Cliente	Prezzo	Somma
3	Valore medio di unità venduta per prodotto, cliente e data	Prodotto, Data, Testata Ordine, Riga Ordine, Cliente	Prezzo, Unità venduta	Media, Somma

Tabella 4.1: Tabella della specifica dei requisiti di analisi per Vendita

Si descrive il fatto *Vendita*, specificando il suo significato, le dimensioni e le misure preliminari interessate. La granularità del fatto determina la

dimensione del Data Mart e il tipo di analisi che si possono effettuare sui dati.

Descrizione	**Dimensioni**	**Misure**
Riga dell'ordine riguardo un prodotto venduto ad un cliente	Prodotto, Data, Testata Ordine, Riga Ordine, Cliente	Prezzo, Unità venduta

Tabella 4.2: Tabella del fatto Vendita

Le dimensioni

Si descrivono le dimensioni specificando per ognuna di esse il nome, una descrizione e la granularità.

Nome	**Descrizione**	**Granularità**
Data	Dimensione temporale che identifica il verificarsi di un fatto	Un giorno
Prodotto	Il prodotto finito incluso nell'ordine di vendita	Il singolo prodotto
Cliente	Il cliente che acquista un prodotto	Il cliente
Testata Ordine	Rappresenta l'identificativo dell'ordine di vendita che raggruppa più righe	L'ordine di vendita
Riga Ordine	Rappresenta la riga dell'ordine di vendita	La riga d'ordine

Tabella 4.3: Tabella delle dimensioni per Vendita

Sono state descritte le dimensioni, ovvero le coordinate di analisi per le aggregazioni di misure dei fatti. In generale, per rendere le analisi più dettagliate è necessario descrivere ogni dimensione con degli attributi utili per le analisi che si vogliono fare e tali che per ogni loro valore sia individuabile un sottoinsieme dei fatti sul quale sia interessante aggregare qualche misura. Di seguito è riportata la tabella riassuntiva delle caratteristiche dell'attributo dimensionale relativo al Prodotto e al Cliente.

Attributo	Descrizione
SKU	Codice univoco del prodotto
Collection	Linea di riferimento del prodotto (ad esempio Sun, Optical ecc.)
Material type	Materiale del prodotto
Brand	Sottocategoria del prodotto (Classic, Junior, Woman ecc.)
Brand category	Categoria del prodotto (ad esempio Lifestyle, Luxury, Premium Fashion ecc.)

Tabella 4.4: Tabella degli attributi dimensionali relativi al Prodotto

Attributo	Descrizione
User ID	Identificatore univoco del cliente
Città	Città di residenza del cliente
Nazione	Nazione di residenza del cliente
Tipo	Tipologia di cliente, registrato o cliente occasionale

Tabella 4.5: Tabella degli attributi dimensionali relativi al Cliente

Le gerarchie dimensionali In seguito vengono evidenziate le relazioni fra gli attributi indicando la dimensione di appartenenza, breve descrizione e tipologia. Come descritto in [Albano 13] una gerarchia fra attributi dimensionali può essere di uno dei seguenti tipi:

- **Bilanciata**. Quando i possibili livelli sono in numero predefinito e i valori degli attributi che ne fanno parte sono sempre definiti.
- **Incompleta**. Quando i possibili livelli sono in numero predefinito, mai valori degli attributi che ne fanno parte non sono sempre tutti definiti.
- **Ricorsiva**. Quando i livelli sono in numero variabile.

Dimensione	Descrizione gerarchia	Tipo gerarchia
Data	Giorno→Mese→Anno	bilanciata
Data	Settimana→Anno	bilanciata
Prodotto	SKU→Material type	bilanciata
Prodotto	SKU→Brand→Brand category	bilanciata
Prodotto	SKU→Collection	bilanciata
Cliente	Città→Nazione	bilanciata

Tabella 4.6: Tabella delle gerarchie dimensionali

Le dimensioni che cambiano Si specifica il tipo di strategia da usare per trattare le dimensioni con attributi che possono cambiare nel tempo, specificando per ognuna di esse, oltre al nome, gli attributi che cambiano e il tipo di strategia per trattarli nella progettazione logica e nel caricamento dei dati. Come riportato in [Albano 13], si considerano quattro possibilità, delle quali le prime tre si considerano quando qualche attributo cambia raramente:

- **Tipo 1 (Riscrittura della storia).** Il valore di un attributo dimensionale che cambia va trattato come un valore errato da sostituire con il nuovo valore. Questa rappresenta la soluzione più semplice che comporta la perdita della storia.
- **Tipo 2 (Conservazione della storia).** Si vuole la storia dei valori. Questa rappresenta la soluzione più comune che va discapito dell'aumento dei dati della dimensione.
- **Tipo 3 (Conservazione di una o più versioni della storia).** Si vuole sia la storia dei valori che la data in cui si verifica il cambiamento.
- **Tipo 4 (Dimensioni che cambiano velocemente).** Gli attributi cambiano frequentemente e non vanno trattati con una delle soluzioni precedenti.

Nome	Attributi modificabili	Trattamento modifiche
Data	Si	Tipo 1
Prodotto	Si	Tipo 1
Cliente	No	Tipo 1
Testata ordine	No	Tipo 1
Riga ordine	No	Tipo 1

Tabella 4.7: Tabelle delle dimensioni che cambiano per Vendita

Le misure

Si presenta la scelta delle misure dei due fatti. Questa scelta avviene sulla base delle metriche da calcolare. Queste misure sono inerenti agli aspetti del business aziendale che si vogliono tenere sotto controllo o analizzare. Vediamo nel dettaglio le misure e forniamo qualche informazione aggiuntiva come il tipo di aggregabilità che, come riportato in [Albano 13], specifica quali funzioni di aggregazione sono applicabili quando si raggruppano i dati secondo certe dimensioni. In particolare va specificato di quale dei seguenti tipi è una misura:

- **Additiva.** Se può essere sommata per ogni dimensione.
- **Semi additiva.** Se non può essere sommata per alcune dimensioni, tipicamente quella che rappresenta il tempo.
- **Non additiva.** Se può essere solo contata o soggetta a media.

Misura	Descrizione	Aggregabilità	Derivata
Unità vendute		additiva	no
Prezzo		additiva	no

Tabella 4.8: Tabelle delle misure del fatto Vendita

4.2.2 Progettazione concettuale inziale dei data mart

Gli attributi citati nelle analisi dei dati suggeriscono come possibile schema concettuale iniziale quello mostrato in figura. Lo schema riporta le misure all'interno della tabella, mentre esternamente sono collocate le relative dimensioni. Questa rappresentazione mette in evidenza importanti dettagli descritti in [Albano 13]:

- **Gerarchie bilanciate**. Si verificano quando i possibili livelli sono in numero predefinito e i valori degli attributi che ne fanno parte sono sempre definiti. Per esempio, gli attributi Day, Month e Year della Data fanno parte di una gerarchia bilanciata con tre livelli;
- **Attributi o dimensioni opzionali**. si verifica quando il valore di una dimensione o di un attributo può essere opzionale e sono modellati con archi tagliati;
- **Attributi descrittivi**. Le dimensioni e gli attributi dimensionali rappresentati con archi che terminano con un circoletto possono essere usati nelle operazioni di analisi per esprimere restrizioni sui loro valori e per fare raggruppamenti o aggregazioni. Quando invece si vogliono rappresentare attributi dimensionali, o attributi dei fatti diversi dalle misure, che non vanno usati nelle operazioni di analisi per fare raggruppamenti, essi vengono detti descrittivi e si rappresentano con archi privi del circoletto.

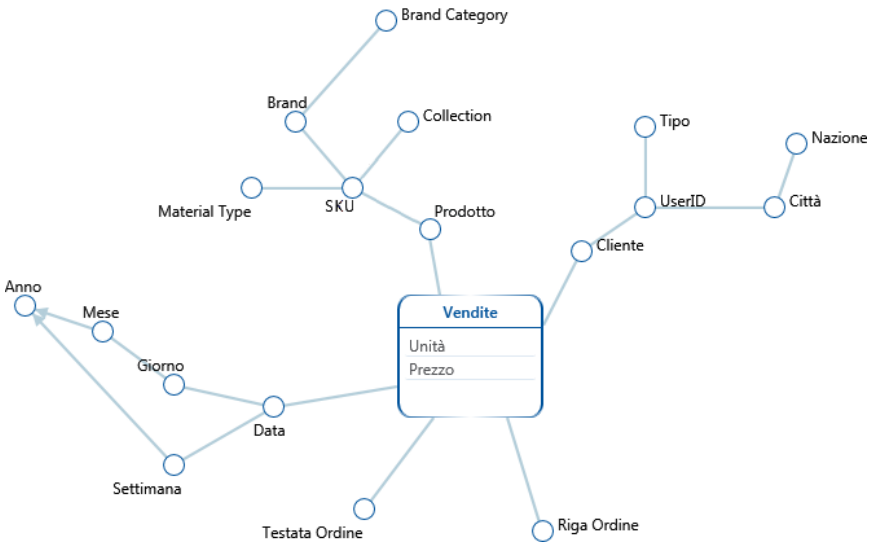


Figura 4.1: Schema concettuale del data mart relativo a Vendita

4.3 Processo di Navigazione del cliente

4.3.1 Specifica dei requisiti del fatto

Di seguito si evidenziano le dimensioni, gli attributi e le misure utilizzate in ogni requisito di analisi inizialmente raccolto relative al flusso di dati Web.

N	Requisito	Dimensioni	Misure	Aggregazioni
1	Totale delle visite per prodotto, per chiave di ricerca, canale di marketing, per cliente, per dispositivo e per data	Data, Prodotto, Canale di marketing, Cliente, Parola chiave, Dispositivo	Visite, Visitatori Unici	Somma
2	Totale dei visitatori unici per prodotto, per chiave di ricerca, canale di marketing, per cliente, per dispositivo e per data	Data, Prodotto, Canale di marketing, Cliente, Parola chiave, Dispositivo	Visite, Visitatori Unici	Somma
3	Quantità media delle pagine visualizzate per prodotto, chiave di ricerca, canale di marketing, cliente, dispositivo e data	Data, Prodotto, Canale di marketing, Cliente, Parola chiave, Dispositivo	Pagine visualizzate per Visita	Media
4	Conversion rate per prodotto, chiave di ricerca, canale di marketing, cliente, dispositivo e data	Data, Prodotto, Canale di marketing, Cliente, Parola chiave, Dispositivo	Unità vendute, Visite, Visitatori Unici	Media
5	Bounce rate per prodotto, chiave di ricerca, canale di marketing, cliente, dispositivo e data	Data, Prodotto, Canale di marketing, Cliente, Parola chiave, Dispositivo	Bounce Rate	Media

Tabella 4.9: Tabella della specifica dei requisiti di analisi per Clickstream

Si descrive il fatto Clickstream, specificando il suo significato, le dimensioni e le misure preliminari interessate. La granularità del fatto determina la dimensione del Data Mart e il tipo di analisi che si possono effettuare sui dati.

Descrizione	**Dimensioni**	**Misure**
Un fatto riguarda il dettaglio relativo alle attività di un utente all'interno del sito Web di commercio elettronico	Data, Prodotto, Cliente, Parola chiave di ricerca, Dispositivo, Canale di marketing	Visite, Visitatori Unici, Unità venduta, Bounce Rate

Tabella 4.10: Tabella del fatto Clickstream

Le dimensioni

Si descrivono le dimensioni specificando per ognuna di esse il nome, una descrizione e la granularità.

Nome	**Descrizione**	**Granularità**
Dispositivo	Rappresenta la tipologia di dispositivo utilizzato dall'utente per accedere al sito Web	Una visita
Parola chiave di ricerca	Parola utilizzata dagli utenti per cercare il sito nei motori di ricerca	Una visita
Canale di marketing	Provenienza del traffico dei visitatori (ricerche fisiologiche, ricerche a pagamento ecc.)	Una visita

Tabella 4.11: Tabella delle dimensioni per Clickstream

Sono descritte le dimensioni, ovvero le coordinate di analisi per le aggregazioni di misure dei fatti. Per le descrizioni relative a Prodotto, Data e Cliente sono valide quelle esposte nella Tabelle delle dimensioni per Vendita.

Le dimensioni che cambiano Si specifica il tipo di strategia da usare per trattare le dimensioni con attributi che possono cambiare nel tempo, specificando per ognuna di esse, oltre al nome, gli attributi che cambiano e il tipo di strategia per trattarli nella progettazione logica e nel caricamento dei dati.

Nome	Attributi modificabili	Trattamento modifiche
Dispositivo	No	Tipo 1
Parola chiave di ricerca	No	Tipo 1
Canale di marketing	No	Tipo 1

Tabella 4.12: Tabelle delle dimensioni che cambiano per Clickstream

Le misure

Si presenta la scelta delle misure dei due fatti. Questa scelta avviene sulla base delle metriche da calcolare. Queste misure sono inerenti agli aspetti del business aziendale che si vogliono tenere sotto controllo o analizzare. Vediamo nel dettaglio le misure e forniamo qualche informazione aggiuntiva come il tipo di aggregabilità.

Misura	Descrizione	Aggregabilità	Derivata
Visita		additiva	no
Visitatori Unici		additiva	no
Bounce Rate		non additiva	si
Pagine visualizzate per visita		non additiva	si

Tabella 4.13: Tabelle delle misure del fatto Clickstream

4.3.2 Progettazione concettuale inziale dei data mart

Gli attributi citati nelle analisi dei dati suggeriscono come possibile schema concettuale iniziale quello mostrato in figura.

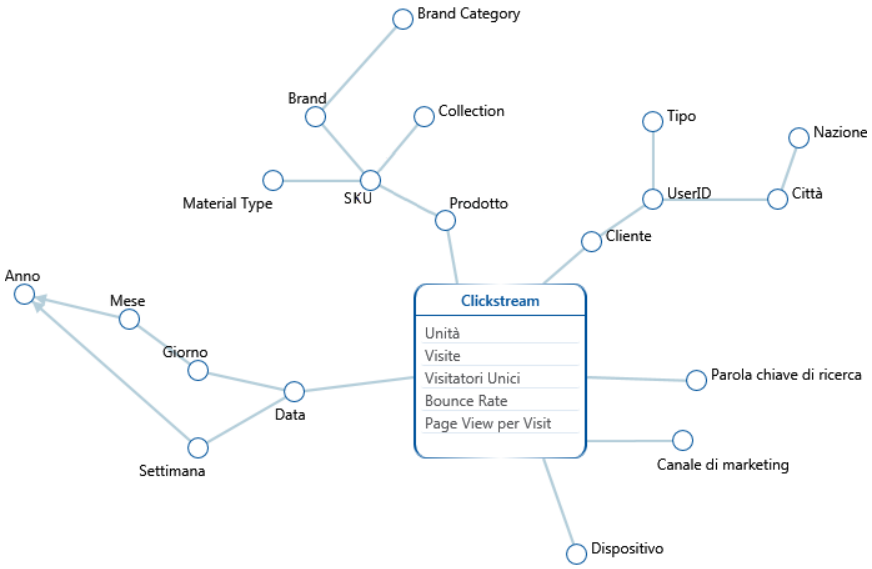


Figura 4.2: Schema concettuale del data mart relativo a Clickstream

4.4 Tabelle riepilogative

Si elencano le dimensioni e le misure individuate specificando in quali fatti si usano. Questo riepilogo è utile per evidenziare quali informazioni sono in comune a processi diversi e quindi dovrebbero avere interpretazione e rappresentazione unica, per essere poi condivise nel Data Warehouse.

Misura	**Clickstream**	**Vendita**
Visite	X	
Visitatori Unici	X	
Unità vendute	X	X
Bounce Rate	X	
Pagine visualizzate per visita	X	
Prezzo		X

Tabella 4.14: Tabella riepilogativa delle misure

Dimensione	**Clickstream**	**Vendita**
Data	X	X
Prodotto	X	X
Cliente	X	X
Parole chiave di ricerca	X	
Dispositivo	X	
Canale di marketing	X	

Tabella 4.15: Tabella riepilogativa delle dimensioni

Capitolo 5

Progettazione concettuale finale e logica dei data mart e del Data Warehouse

Terminata la realizzazione concettuale del data warehouse si può passare alla sua progettazione logica, la quale consiste nella creazione dello schema relazionale del data warehouse trasformando ogni data mart in uno schema relazionale. È necessario uno studio della struttura dei sistemi sorgenti e dei dati in essa contenuti affinché sia possibile una corretta e coerente estrazione ed integrazione. In questo capitolo si descrivono le fonti dati utilizzati nella fase di ETL e successivamente si illustra la progettazione concettuale finale e logica del data warehouse.

5.1 Introduzione ai sistemi sorgenti

5.1.1 Omniture Site Catalyst

I dati relativi al sito di commercio elettronico in analisi, sono forniti dallo strumento di web analytics *Omniture Site Catalyst*. Come riportato in Figura 5.2, lo strumento offre la possibilità di effettuare estrazioni personalizzabili dei dati, forniti in formato testo tramite protocollo di trasferimento file (FTP), e schedulare le estrazioni per la consegna automatica.

Al fine di eseguire una procedura automatica di estrazione dati per il caricamento giornaliero del data warehouse è stata configurata una schedulazione contenente le informazioni di un singolo giorno (il precedente alla data di consegna). Di seguito vengono elencate le informazioni selezionate per l'estrazione creata con lo strumento *Omniture Request Tool*.

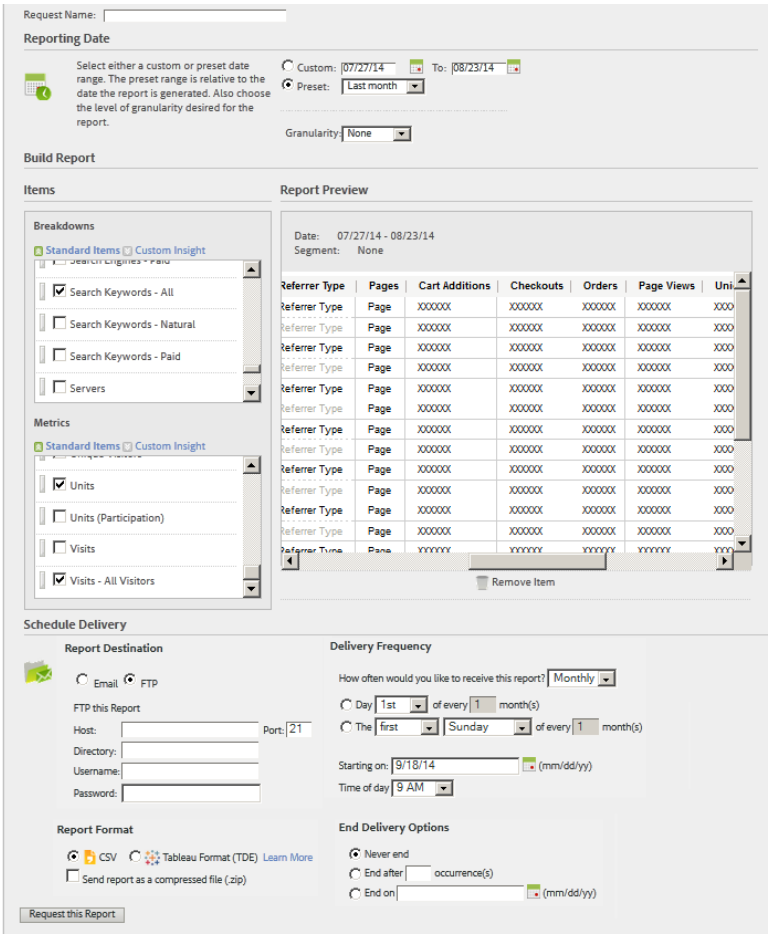


Figura 5.1: Omniture Request Tool

Cookie L'informazione relativa ai cookie è utilizzata per monitorare il comportamento dei visitatori. L'identificatore univoco contenuto in ognuno di essi rappresenta il singolo visitatore (si veda Paragrafo 2.3.1) ed è contenuto nell'attributo VISITOR_ID della tabella rappresentata in Figura 5.2.

Canale di marketing L'informazione si riferisce all'ultimo canale di marketing da cui viene coinvolto il visitatore prima di generare un ordine o qualsiasi operazione all'interno del carrello virtuale (si veda Paragrafo 3.2.2). L'informazione relativa al Canale di marketing è contenuta nell'attributo LT_CHANNEL_NAME.

Pagina L'informazione contiene l'URL della pagina web visualizzata dall'utente, non sono rappresentate le pagine temporanee memorizzate nella

cache del web server.

Mandante o Referrer Rappresenta l'indirizzo URL di una risorsa del web in cui si trovavano i visitatori prima di raggiungere il sito.

Parole chiave di ricerca Le parole chiave di ricerca sono contenute nell'attributo KEYWORD_NAME e rappresentano le *corrispondenze generiche* dei termini inseriti nei motori di ricerca dai visitatori. Per corrispondenza generica si intende l'associazione di un testo di ricerca ad uno o più termini generici, indipendentemente dall'ordine in cui tali termini sono stati inseriti all'interno del testo. L'associazione è garantita anche nel caso in cui ci siano errori di ortografia, forme al plurale, acronimi ecc.

Area Geografica Informazioni relative alla posizione geografica, ossia lo Stato e la città di provenienza delle visite.

Carrello virtuale Informazioni riguardanti il numero di prodotti inseriti nel carrello (Cart Addition), il numero di visualizzazioni del suo contenuto (Cart Views) e il numero di volte in cui il visitatore ha effettuato il checkout dell'ordine. Con checkout si intende la fase antecedente alla conclusione dell'acquisto.

Ordine di vendita Nel caso in cui l'utente abbia effettuato uno o più acquisti, vengono valorizzati le seguenti informazioni: gli identificatori univoci dei prodotti acquistati, come il codice articolo gestito a magazzino (SKU) e la codifica universale (UPC); la quantità di articoli ordinati e il numero di ordini generati, informazioni contenute rispettivamente negli attributi UNITS e ORDERS (si veda Tabella 5.2).

Indicatori derivati

- Il *Conversion Rate* è calcolato dal rapporto tra numero di ordini (ORDERS) e numero di visitatori unici (VISITOR_ID);
- il *Bounce Rate* è disponibile in forma pre-calcolata;
- *Page View per Visit* è disponibile in forma pre-calcolata;

5.1.2 IBM WebSphere Commerce

IBM WebSphere Commerce o WCS è una architettura logica di supporto su cui il sito di commercio elettronico è stato realizzato. Il framework si basa su un'architettura a tre livelli:

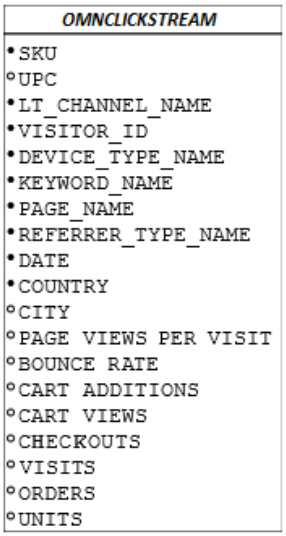


Figura 5.2: Omniture Data

- Un database (IBM DB2, Oracle ecc.);
- Una application server (Java EE);
- Un web server (IBM HTTP Server, Microsoft Information Services ecc..).

IBM WebSphere Commerce offre la possibilità di accedere direttamente al database, all'interno del quale sono contenute sia le tabelle attinenti agli ordini generati, sia le tabelle del catalogo prodotti. Non sono state riscontrate problematiche riguardo la possibilità di schedulare giornalmente una estrazione totale o parziale dei dati contenuti per il processo di ETL.

Di seguito si elencano le informazioni disponibili ed utilizzate per la progettazione finale del data mart relativo alle vendite.

Utente Rappresenta le informazioni relative agli utenti registrati. Per utente registrato si intende colui che si iscrive fornendo informazioni personali per ottenere accesso all'utilizzo di alcuni servizi che non sono messi a disposizione della generalità degli utenti. I visitatori possono generare ordini anche se non registrati. Le informazioni inerenti alla generalità degli utenti sono contenute nella tabella USERS. Nella tabella ADDRESS invece, sono contenute le informazioni relative all'indirizzo quali, la città, lo Stato e l'indirizzo di fatturazione e di spedizione dell'ordine.

Catalogo Il catalogo contiene la relazione tra gli articoli presenti nell'indice di catalogo e gli ordini creati dagli utenti. È rappresentato nel database dalla tabella CATENRTY.

Indice di catalogo L'indice di catalogo tiene traccia dei prodotti esistenti nel negozio online e dei relativi attributi, quali ad esempio nome, codice, descrizione, uno o più prezzi, immagini e altri dettagli. L'indice di catalogo è rappresentato dalla tabella CATENTRYATTR che si lega con la tabella ATTR per recuperare gli attributi di un prodotto.

Prodotto e Articolo Un prodotto è il modello per un insieme di articoli definiti dagli stessi attributi indicati nell'indice di catalogo. Ad esempio, un occhiale rappresenta un prodotto generico del catalogo. Un articolo, invece, è un prodotto a cui sono stati valorizzati gli attributi (riferendoci all'esempio dell'occhiale: colore e taglia). Ogni combinazione possibile attributo-valore definisce un articolo.

Valori di attributi I valori di attributo sono il dominio di valori che può avere un attributo, quali ad esempio un determinato colore (blu o giallo) o la taglia (piccola, media o grande). Nel database queste informazioni sono contenute nella tabella ATTRVAL.

Ordini L'ordine descrive la transazione commerciale relativa all'acquisto di uno o più articoli da parte dell'utente. Un ordine di vendita è rappresentato dalla testata e da una riga d'ordine per ogni articolo acquistato. Un ordine deve avere almeno un articolo.

Ad ogni testata d'ordine è associata:

- indirizzo di spedizione e fatturazione;
- spese di spedizione e tasse;
- modalità di spedizione;
- altro.

Per ogni articolo ordine è associato:

- un centro di evasione ordini;
- sconto, prezzo totale e quantità;
- altro.

Nel database le tabelle ORDER e ORDERITEMS, rappresentano rispettivamente le testate e le righe d'ordine.

Canale di vendita Rappresenta il canale di vendita su cui è stato effettuato l'ordine (Italy, USA, Canada ecc.), informazioni presenti nella tabella BUSCHN.

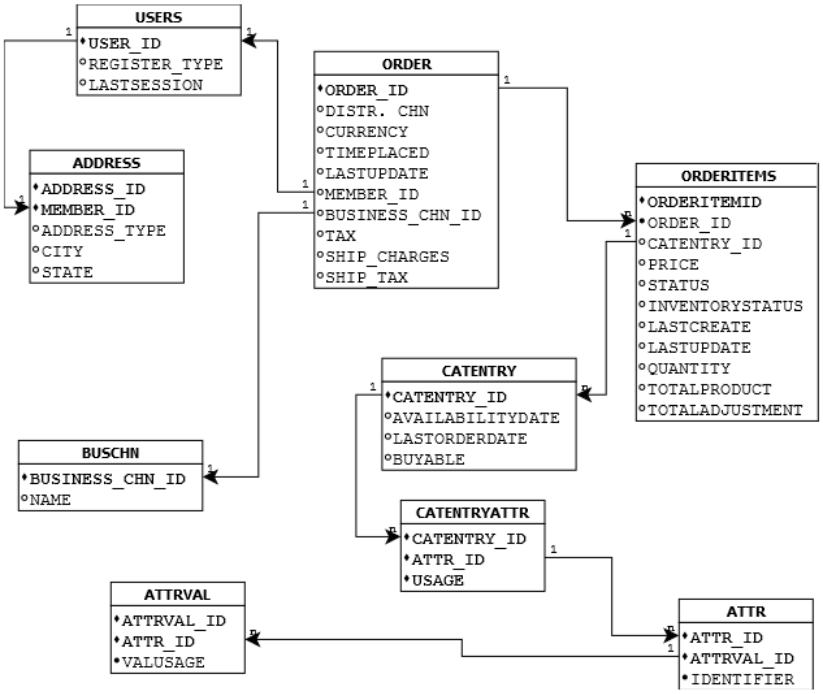


Figura 5.3: WebSphere Commerce Suite

5.1.3 Sistema gestionale SAP R/3

Il sistema SAP R/3 gestisce tutte le informazioni rilevanti dell'azienda, permettendo l'accesso ai dati gestionali a tutte le funzioni dell'azienda. Il sistema è strutturato in vari moduli a seconda della funzione aziendale a cui sono rivolti. L'anagrafica prodotti è gestita dal modulo SAP relativo al controllo di magazzino, il cui scopo è quello di controllare i movimenti e il deposito di materiali e/o prodotti nel magazzino e di processarne le transizioni (spedizione, ricezione, riordino e raccolta). Alla luce di tale informazione è risultato conveniente individuare nel database del relativo modulo, le tabelle riguardanti il prodotto. Tali tabelle (si veda in Figura 5.4) hanno come codice univoco lo stock keeping unit (SKU), che essendo comune alle sorgenti IBM WebSphere Commerce e Site Catalyst, permette di integrare attributi dimensionali non presenti nei due sistemi sorgente.

Gli attributi in questione sono:

- BRAND sottocategoria del prodotto;

- MAT_TYPE materiale del prodotto;
- COLLETION linea di riferimento del prodotto;
- BRAND_CATEGORY categoria del prodotto.

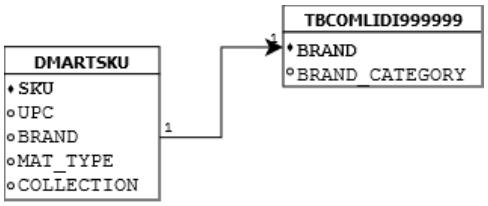


Figura 5.4: Anagrafica Prodotto

Le informazioni sulla consegna e fatturazione dell'ordine sono presenti nel modulo SAP relativo alla fatturazione (si veda Figura 5.5). In particolare:

Ordini La tabella FD3ORDINE contiene l'insieme degli ordini ed è legata alla tabella ORDERS di WCS tramite l'attributo ORDERID. Gli ordini online sono identificati dal tipo ordine *ZCP1* (attributo TYPE).

Consegne La tabella FD3CONS contiene l'insieme delle consegne ed è legata alla tabella degli ordini tramite il codice ordine. Quando la consegna viene generata è nello stato *pending*; quando viene presa in carico viene creata una pick list, una lista che specifica l'attività di prelievo a magazzino degli articoli nella tipologia e quantità previste; quando viene emessa la bolla si definisce la data di consegna ed infine viene emessa la fattura.

Fatture La tabella FD3FATTURA contiene l'insieme delle fatture contabilizzate. È legata alla tabella degli ordini tramite il codice ordine. Uno stesso ordine può essere diviso in più fatture. Il valore dell'importo è presente in tre valute (valuta di fatturazione, valuta della società che emette la fattura, valuta in Euro).

Ordini Cancellati e Resi I resi sono gestiti con ordini appositi che generano consegne fittizie e fatture con valori negativi. Essi sono legati al cliente e al prodotto e non ad un particolare ordine o fattura. Infine, gli ordini cancellati sono aggiornati direttamente sulla fattura interessata, assegnando al campo TIMEPLACED un valore fittizio.

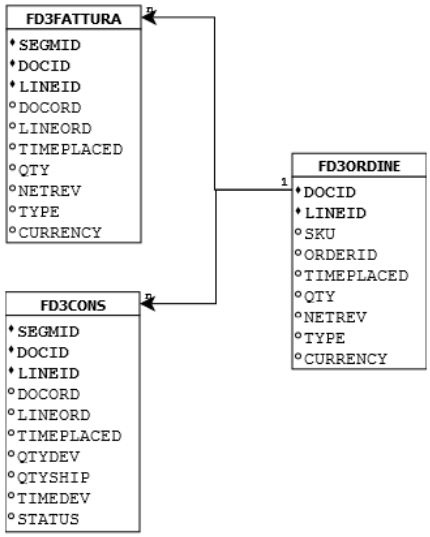


Figura 5.5: Gestione fatture

5.2 Classificazione delle tabelle interessanti

In questa sezione si esaminano gli schemi relazionali per decidere quali tabelle e attributi sono interessanti secondo quanto stabilito durante la raccolta dei requisiti (si veda il Capitolo 3). L'obiettivo è di garantire che ai fatti modellati corrispondano dei dati operazionali e che da questi si possano estrarre altre informazioni utili, arricchendo così gli schemi dei data mart iniziali.

Analisi dei dati operazionali In questo passo, si analizzano le sorgenti dati descritte in precedenza per compiere due azioni:

- uniformare la terminologia e le unità di misura delle grandezze numeriche che devono avere lo stesso riferimento temporale;
- eliminare le tabelle e gli attributi ritenuti non interessanti ai fini dell'analisi dei dati.

Prima di progettare i data mart a partire dalle basi di dati, si evidenziano le tabelle e gli attributi utili per la costruzione degli schemi finali, classificando le tabelle in tre categorie:

- **Entità evento.** Sono le tabelle che rappresentano eventi potenzialmente interessanti per i processi aziendali di analisi dei dati. Le tabelle classificate come entità evento sono: FD3ORDINE, FD3FATTURA, FD3CONS, ORDERS, ORDERITEMS e OMNCLICKSTREAM.

- **Entità componente.** Sono le tabelle in relazione con un'entità evento con un'associazione 1:n. Le tabelle classificate come entità componente sono: DMARTSKU, USERS, BUSCHN e CATENTRY.
- **Entità di classificazione.** Sono le tabelle in relazione con un'entità componente con una catena di associazioni 1:n. Le tabelle classificate come entità di classificazione sono: TBCOMLIDI999999, CATENTRYATTR, ATTRVAL, ATTR e ADDRESS.

5.3 Integrazione dei dati operazionali negli schemi concettuali dei data mart iniziali

L'analisi delle sorgenti informative è risultata indispensabile per passare dagli schemi dei data mart iniziali a quelli finali. Considerando le informazioni aggiuntive che si sono integrate dal livello operazionale, sono emerse nuove opportunità di analisi non rilevabili durante la prima raccolta dei requisiti.

- Verranno integrati nel data mart delle vendite le informazioni inerenti:
  - al canale di vendita e al tipo di dispositivo presso il quale è stato emesso un ordine (tabella BUSCHN);
  - al prezzo netto e al prezzo lordo dell'articolo (SAP R3 modulo di fatturazione);
  - allo stato di un ordine (si veda Tabella 5.1).
- Verranno integrati nel data mart del processo di navigazione le informazioni contenute nella tabella OMNCLICKSTREAM, quali:
  - *Referrer Type*, per comprendere meglio le dinamiche di posizionamento del sito sui motori di ricerca e sul web in generale;
  - carrello virtuale (*Cart Addition, Cart View e Checkout*), per l'analisi sul processo di acquisto dei potenziali clienti;
  - tipo di dispositivo utilizzato dal visitatore, allo scopo di discriminare gli utenti che usufruiscono della piattaforma *mobile*;
  - pagina web visualizzata dall'utente, per valutare quanto gli utenti apprezzino i contenuti del sito;
  - numero ordini effettuati dal visitatore e Conversion Rate in forma pre-calcolata.

Sulla base delle nuove informazioni, i data mart subiscono delle variazioni. La granularità dei due fatti risultano uguali a quanto esposto nelle analisi dei requisiti (si veda Capitolo 4).

Stato	Sorgente	Descrizione
P	WCS	Ordine generato dal visitatore ma non ancora confermato.
A	WCS	Pagamento dell'ordine in attesa di approvazione.
M	WCS	Pagamento autorizzato.
U	WCS	Ordine in attesa di essere inviato al sistema di gestione magazzino.
V	WCS, SAP	Ordine in fase di trattamento.
D	WCS, SAP	Ordine spedito e fatturato.
B	WCS, SAP	Reso.
K	WCS, SAP	L'invio dell'ordine al sistema gestionale è fallito.

Tabella 5.1: Stato di un ordine

5.4 Data Mart delle Vendite

Dopo aver analizzato le sorgenti dati viene definita la dimensionalità potenziale del fatto degli ordini che ammonta a circa 1.000.000 di record. Il sistema WCS elabora dai 400 ai 500 ordini al giorno.

5.4.1 Progettazione concettuale finale del data mart

In Figura 5.6 viene mostrato lo schema concettuale finale del data mart relativo agli ordini di vendita, con l'integrazione delle dimensioni Stato Ordine, Dispositivo, Canale di Vendita e la misura Prezzo Netto.

5.4.2 Progettazione logica del data mart

Come proposto da [Albano 13], in questa fase si passa dagli schemi finali dei data mart allo schema relazionale del data warehouse, trasformando gli schemi concettuali finali dei data mart nei due schemi relazionali. Per le due tabelle dei fatti viene creata una tabella relazionale, con una chiave surrogata e tutte le misure come attributi. Successivamente viene creata una tabella relazionale per ogni dimensione presente nello schema concettuale finale. Anche queste tabelle contengono una chiave primaria surrogata e tutti gli attributi presenti nella rispettiva dimensione. In figura 5.7 viene mostrato lo schema logico del data mart delle vendite. La dimensione sullo stato dell'ordine diventa attributo della tabella del fatto. Le tabelle del sistema gestionale SAP/R3 sono utilizzate per legare la metrica relativa al prezzo netto alla tabella dei fatti e per inglobare gli stati dell'ordine che non erano compresi in WCS.

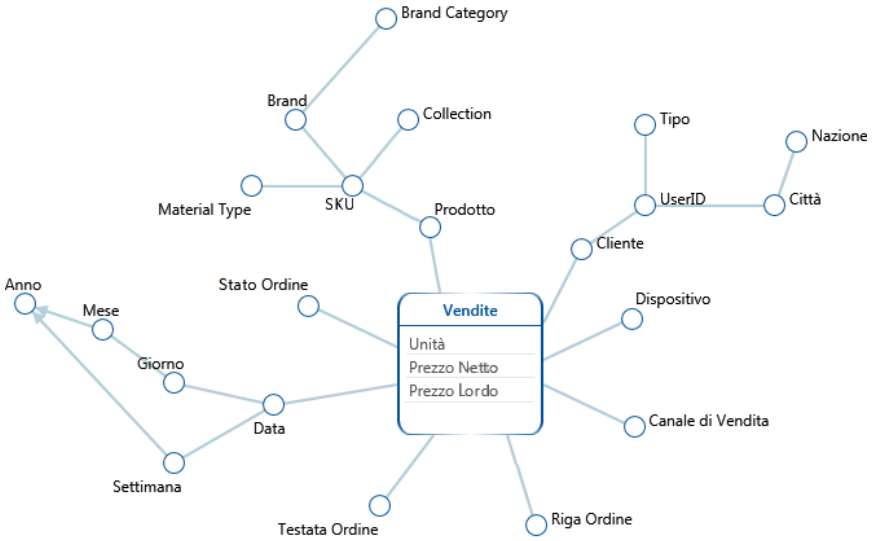


Figura 5.6: Schema concettuale finale del data mart delle vendite

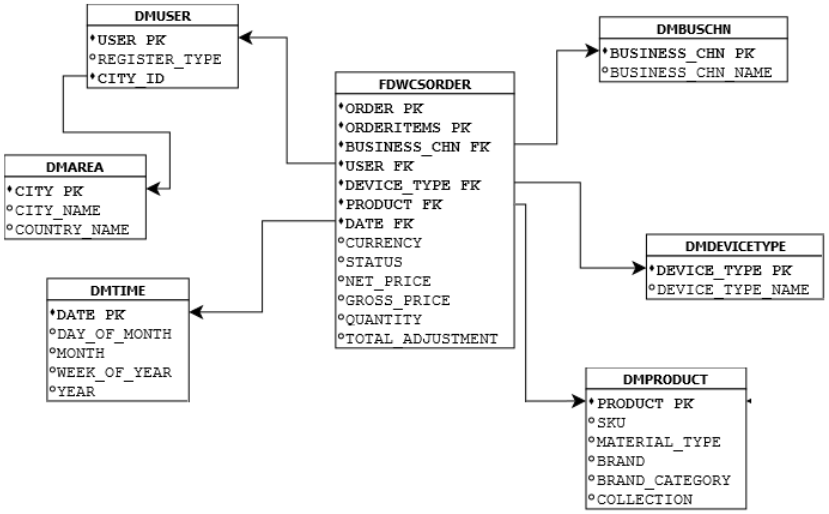


Figura 5.7: Schema logico del data mart delle vendite

5.5 Data Mart del processo di navigazione

Dopo aver analizzato le sorgenti dati viene definita la dimensionalità potenziale del fatto inerente il processo di navigazione che ammonta a circa 50.000.000 record. Questa informazione potrebbe influenzare le scelte di progettazione logica del data mart.

5.5.1 Progettazione concettuale finale del data mart

In Figura 5.5.1 viene mostrato lo schema concettuale finale del data mart relativo al processo di navigazione, con l'integrazione delle dimensioni Referrer Type e Pagina, e le misure Cart Addition, Cart View, Numero Ordini, Checkout e Conversion rate.

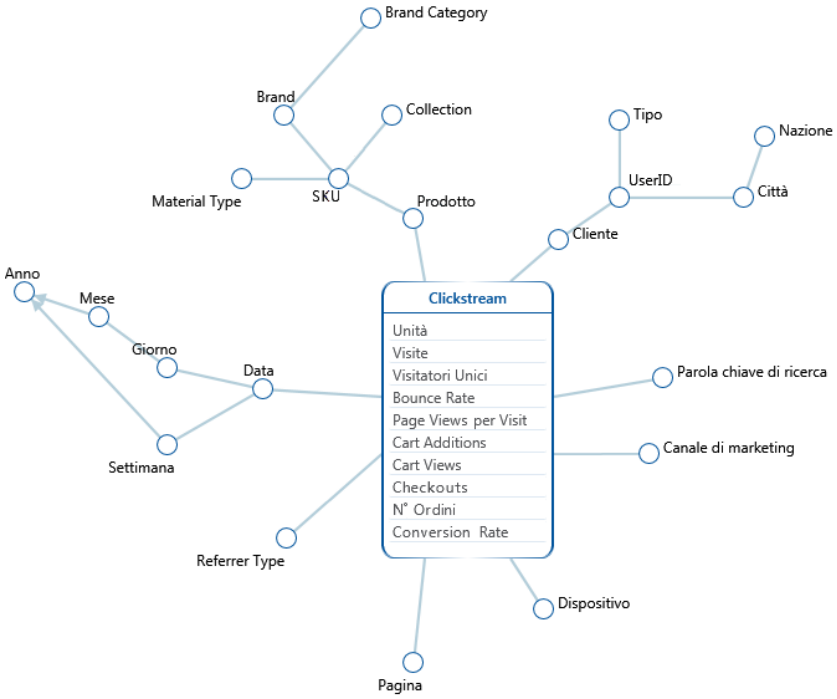


Figura 5.8: Schema concettuale del data mart relativo alla navigazione dell'utente

5.5.2 Progettazione logica del data mart

In figura 5.9 viene mostrato lo schema logico del data mart. Da notare che le tabelle relazionali delle dimensioni Prodotto, Utente, Dispositivo e Tempo presenti, sono condivise con il data mart delle vendite.

5.6 Progettazione logica del Data Warehouse

Le dimensioni in comune sono rappresentate da un'unica tabella, non si sono riscontrati problemi o necessità di creare viste per rinominare o escludere alcuni attributi. In Figura 5.10 si presenta lo schema logico complessivo del data warehouse con le dimensioni in comune e condivise dai due data mart.

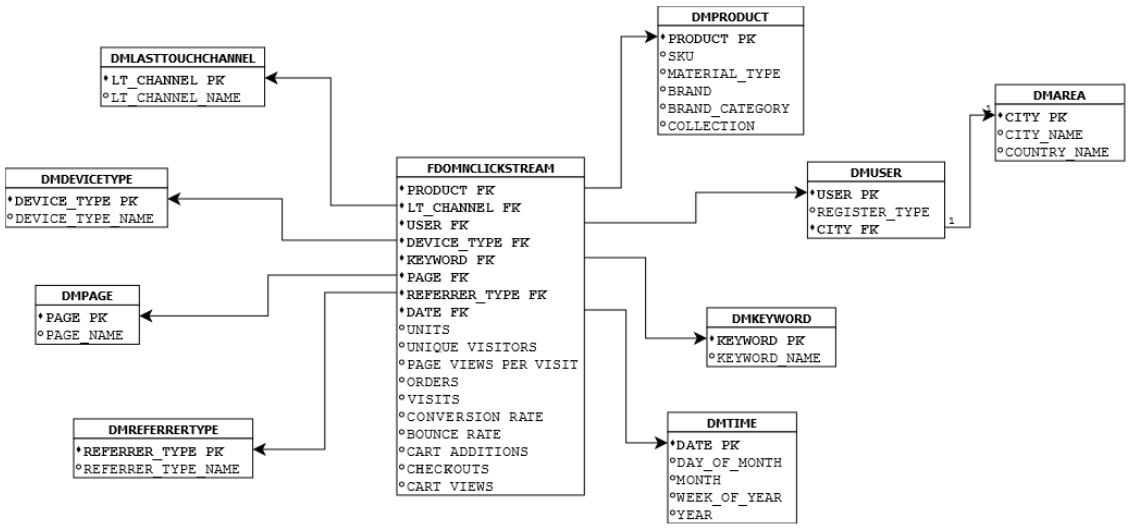


Figura 5.9: Schema logico del data mart relativo alla navigazione dell'utente

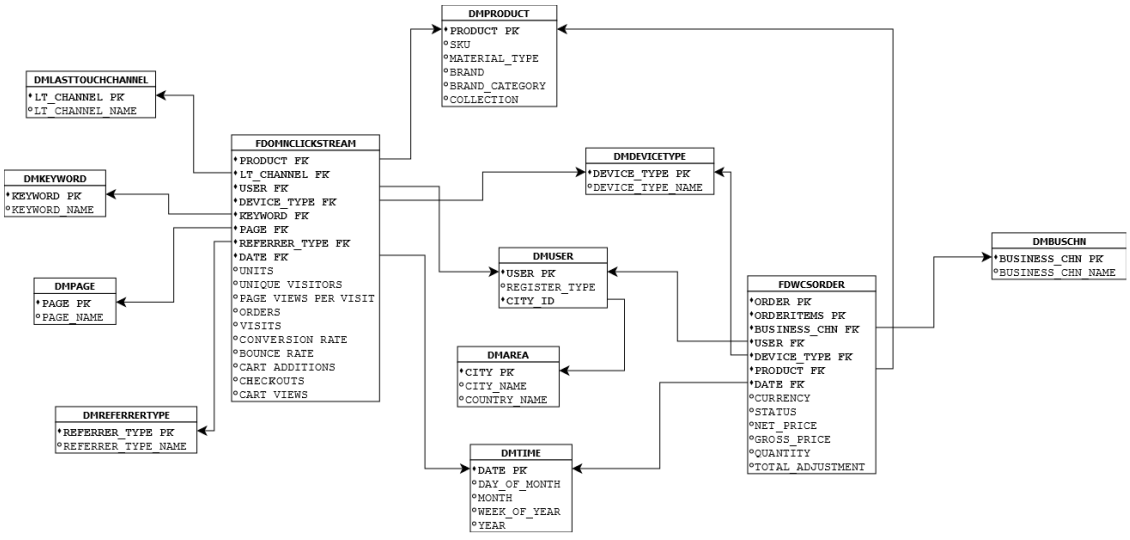


Figura 5.10: Schema logico del Data Warehouse

Capitolo 6

Ambiente di sviluppo

In questo capitolo si presentano gli strumenti utilizzati per lo svolgimento dell'intero progetto di data warehousing. Sono illustrate le caratteristiche generali del DBMS *IBM Netezza* per la memorizzazione fisica del data warehouse, le componenti principali di *SAP Business Objects Data Services* per l'estrazione, trasformazione e caricamento dei dati e di *QlikView* per la reportistica finale dei dati.

6.1 IBM Netezza 1000

6.1.1 Architettura

Come riportato in [Netezza 09], lo strumento IBM Netezza è un'architettura a parallelismo massivo (MPP), che integra database, server, storage e funzionalità analitiche. Di seguito si riportano i principi su cui si basa la sua architettura:

Elaborazione dei dati vicina alla loro sorgente. La soluzione IBM Netezza utilizza componenti di commodity, i cosiddetti dispositivi *FPGA* (*Field Programmable Gate Array*), per filtrare i dati non necessari al processo a monte del flusso dati. Questo processo di eliminazione dei dati vicina alla sorgente elimina i colli di bottiglia dell'I/O, mantenendo libere le componenti a valle come la CPU, la memoria e la rete, non più costrette a elaborare dati superflui. Il risultato è una visibile accelerazione delle prestazioni complessive.

Piattaforma per Advanced Analytics. Netezza consente di incorporare algoritmi non-SQL complessi negli elementi di elaborazione dei suoi flussi. Questo è reso possibile grazie alla coordinazione delle componenti, come i dispositivi FPGA precedentemente descritti, il processore e la memoria. In

questo modo tali componenti operano simultaneamente, massimizzando l'uso e ottenendo il massimo throughput. La possibilità di eseguire analisi su quantità di dati consistenti, delegando i ritardi e i costi di trasferimento ad un hardware separato, accelera di alcuni ordini di grandezza le prestazioni.

Scalabilità e configurazioni flessibili. IBM Netezza è scalabile da poche centinaia di gigabyte a decine di petabyte di dati. L'architettura del sistema è oltremodo adattabile per soddisfare le esigenze di diversi segmenti del mercato, dal data warehousing al big data analytics. L'impiego di componenti *open-based*, ovvero di componenti altamente configurabili, consentono di modificare facilmente il rapporto disco-processore-memoria in configurazioni orientate maggiormente alle prestazioni oppure alla capacità. La stessa architettura supporta anche sistemi *memory-intensive* che offrono analisi estremamente veloci e in tempo reale per applicazioni critiche.

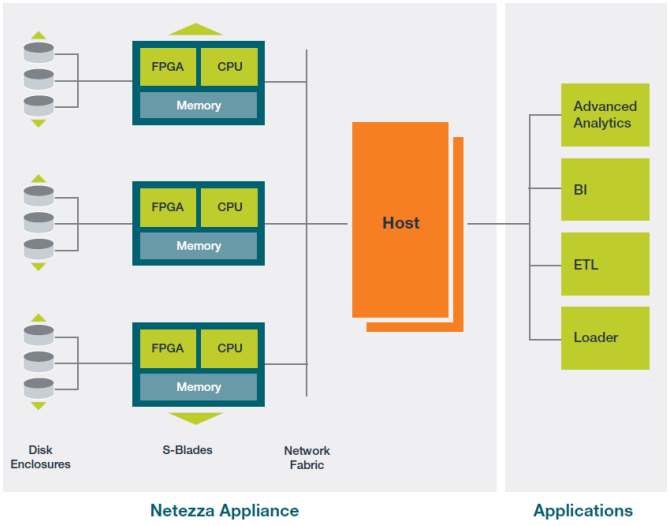


Figura 6.1: Architettura del sistema IBM Netezza

6.1.2 Gli elementi del sistema

In questa sotto sezione si elencano gli elementi principali che caratterizzano l'architettura di IBM Netezza (si veda Figura 6.1).

Host. Gli host sono server Linux ad alte prestazioni impostati per massimizzare la disponibilità delle risorse. L'host: compila le query SQL ottenendone segmenti di codice eseguibili, detti *snippet*; crea dei piani di esecuzione ottimizzati e distribuisce gli *snippet* ai vari nodi *MPP* perché siano eseguiti.

Snippet Blade. Rappresentano i nodi di elaborazione che costituiscono il motore *MPP* dell'appliance Netezza. Ogni *S-Blade* è un server indipendente che contiene CPU, dispositivi *FPGA* e diversi gigabyte di RAM, il tutto bilanciato e concorrente. I core delle CPU sono progettati per eseguire algoritmi complessi su grandi volumi di dati per applicazioni analitiche avanzate.

Storage Array. Contengono dischi ad alta densità e alte prestazioni con protezione *RAID*. Ogni disco contiene una parte dei dati appartenenti ad una tabella del database. Gli alloggiamenti dei dischi sono collegati alle *S-Blade* tramite interconnessioni ad alta velocità, che permettono a tutti i dischi di trasmettere i dati simultaneamente alle *S-Blade*.

Network Fabric. Ogni componente del sistema è interconnessa attraverso una rete. L'architettura applica un protocollo *IP-based* personalizzato, che sfrutta appieno l'ampiezza di banda delle varie sezioni ed elimina i congestionamenti anche in caso di traffico sostenuto. La rete è ottimizzata per consentire scalabilità fino ad un migliaio di nodi.

6.2 SAP Business Object Data Services XI 3.2

Con l'acquisizione di Business Objects, SAP ha saputo sfruttare la sua grande base installata di applicativi gestionali per promuovere l'adozione del prodotto di Data Integration e Business Intelligence presso i propri clienti. Questo ha consentito a SAP di conquistare una buona fetta di mercato in tempi relativamente brevi. Il prodotto offerto da SAP garantisce funzionalità avanzate di modellazione dei dati e di gestione dei metadati in scenari di integrazione differenti. Il prodotto include il supporto alla federazione dei dati (Business Objects Data Federator) e la piattaforma Data Services, che combina le funzionalità di integrazione con quelle per la gestione della qualità del dato. È soprattutto in quest'ultima componente che si nota l'esperienza di Business Objects, che permette al prodotto di SAP di competere con i leader del mercato.

Come riportato in [SAP 14], *SAP Business Objects Data Services* o SAP BODS è un ambiente che fornisce: interfaccia di sviluppo, gestione della connettività dei dati, repository per i metadati e una console di gestione. Una delle funzionalità fondamentali di Data Services è l'estrazione, trasformazione e caricamento di dati provenienti da fonti eterogenee in un database di destinazione o data warehouse.

Data Services è progettato per prestazioni elevate in un ampio spettro di scenari:

- Gli sviluppatori possono integrare Data Services con il sistema gestionale aziendale SAP attraverso l'utilizzo di web services, Java o .NET API.
- Gli utenti finali possono accedere, creare, modificare e interagire con i progetti e i report di Data Services utilizzando strumenti che includono: Designer e Management Console.
- Gli utenti IT possono usare strumenti di gestione descritti in seguito: Central Management Console, Management Console, Server Manager e Repository Manager.

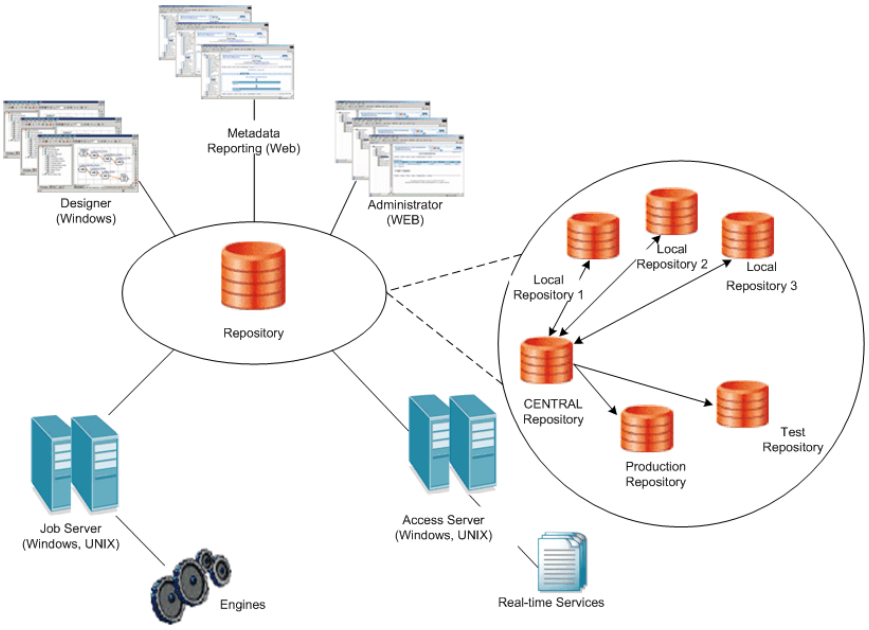


Figura 6.2: Architettura di SAP BO Data Services

6.2.1 Architettura

In questa sezione si descrivono i componenti principali dell'architettura e dei servizi che compongono la suite di SAP Business Objects Data Services. Le informazioni presentate consentono di comprendere gli elementi essenziali per poi essere approfondite nel capitolo successivo.

6.2.2 Componenti principali

Designer Il *Designer* consente di creare, testare ed eseguire flussi dati detti *Job*, per il popolamento del data warehouse. L'interfaccia grafica, consente di creare delle strutture o *Workflow* che regolano: il flusso dati, le

regole di controllo e le logiche di mappatura e trasformazione dei dati (*Dataflow*). Tutte le entità che possono essere definite, modificate e utilizzate nel *Designer* sono chiamate oggetti. Gli oggetti in Data Services rappresentano i metadati. L'interfaccia grafica fornita (si veda Figura 6.3) permette di gestire gli oggetti memorizzati nei repository.

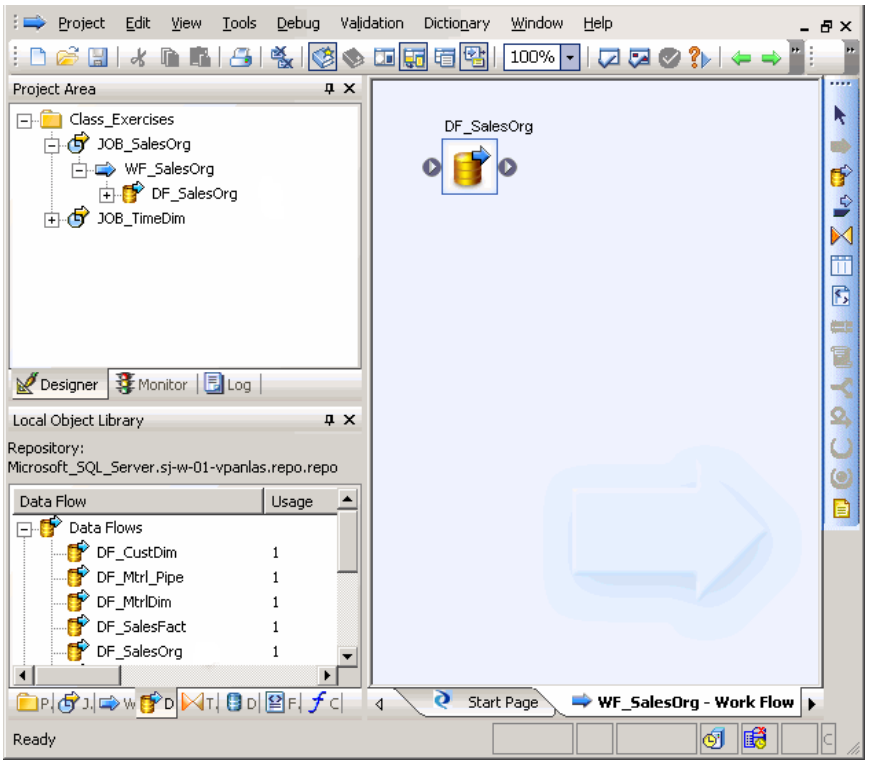


Figura 6.3: Designer di SAP BO Data Services

Repository Un repository è un insieme di tabelle che contengono gli oggetti, metadati e trasformazioni predefinite o create dagli utenti. Ogni repository è memorizzato su un RDBMS registrato nel *Central Management Console*. Ogni repository è associato a uno o più *Job Server*.

Esistono due tipi di repository:

- **Local repository**: usato dal Designer per memorizzare la definizione degli oggetti (project, job, workflow e dataflow) e dei metadati.
- **Central repository**: un componente facoltativo che può essere utilizzato per sostenere lo sviluppo multiutente. Il central repository fornisce una libreria di oggetti condivisi permettendo agli sviluppatori di sincronizzare i propri oggetti da/verso i loro repository locali. Mentre ogni utente lavora su applicazioni in un repository locale unico, il

team utilizza un repository centrale per memorizzare la copia master di tutto il progetto. Il repository centrale conserva tutte le versioni degli oggetti di un'applicazione, in modo da poter tornare a una versione precedente, se necessario.

La sezione *Local Object Library* visibile nel Designer, mostra gli oggetti che sono disponibili nel repository. Esistono due tipologie di oggetti: riusabili e non riusabili. Quelli riusabili, la cui definizione viene salvata all'interno del repository, possono essere richiamati e utilizzati più volte. Se si modifica la definizione dell'oggetto in un posto e si salva, il cambiamento si riflette in tutte le altre sue chiamate. Il *Local Object Repository* mostra quante volte l'oggetto viene richiamato, potendone tener traccia del suo utilizzo.

Alcuni tipici oggetti riusabili sono:

- job, workflow e dataflow;
- tabelle dei datastore e file di testo;
- funzioni.

Gli oggetti non riusabili operano nel contesto in cui sono stati creati. Sono definiti nel singolo Job o Dataflow.

Alcuni oggetti non riusabili sono:

- script;
- annotazioni e commenti;
- oggetti per la gestione di cattura degli errori;
- oggetti per il flusso condizionale dei dati;
- datastore.

Job Server Il Job Server di SAP BODS esegue trasformazioni dati complesse e gestisce estrazioni di dati consistenti dai sistemi ERP o altri sistemi sorgenti. Tutte le operazioni descritte possono essere eseguite in modalità batch o in tempo reale, utilizzando: un sistema distribuito di ottimizzazione delle query, un processo di caching dei dati ed un'elaborazione parallela per fornire scalabilità ed un elevato *throughput* dei dati.

I Job possono essere avviati anche tramite Designer: in questo caso il Designer invoca il Job Server, che richiama il Job dal repository, per poi eseguirlo. È possibile creare gruppi di Job Server per bilanciare il carico di lavoro.

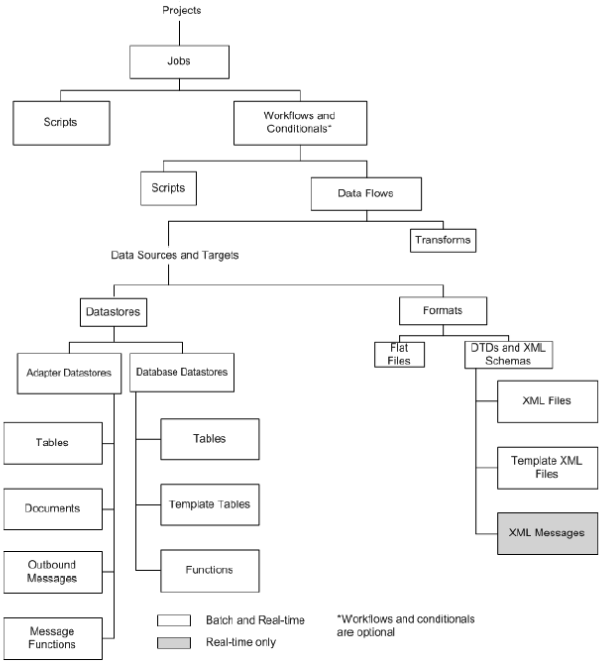


Figura 6.4: Gerarchie di oggetti in *SAP BODS*

Management Console Il *Management Console* di SAP BODS fornisce un pannello di amministrazione gestibile tramite browser, che permette di:

- pianificare, monitorare, ed eseguire *Job*;
- configurare, avviare e terminare servizi real-time;
- configurare Job Server e repository;
- configurare e gestire le connessioni dati;
- gestire gli utenti;
- pubblicare Job via Web Service.

Metadata Reports Applications Sono presenti una serie di strumenti, integrati al Management Console, dedicati al reporting e all'analisi dei metadati, tra i quali:

- Impact and Lineage Analysis
- Operational Dashboards
- Auto Documentation
- Data Validation

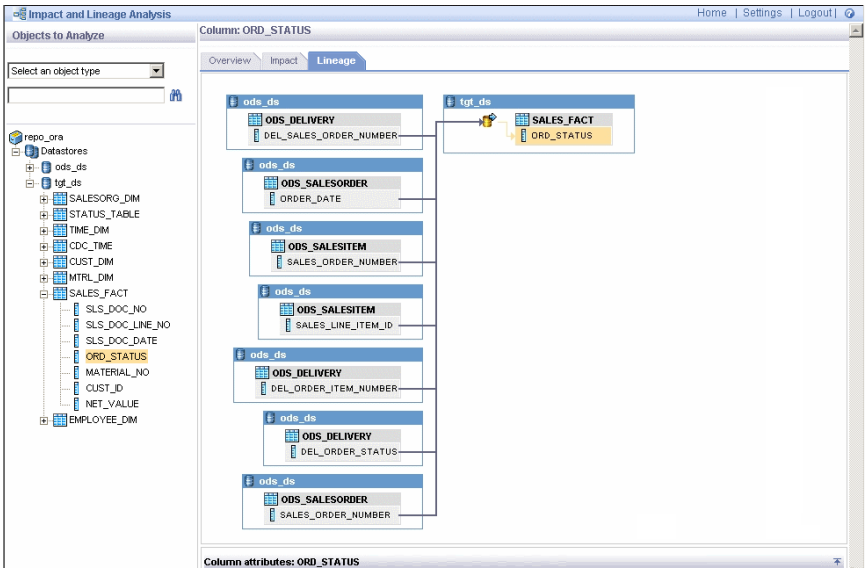


Figura 6.5: Impact and Lineage Analysis

Impact and Lineage Analysis I report di *Impact e Lineage Analysis* possono essere dedicati ad ogni tabella, funzione o relazioni di gerarchia presente nei vari datastore e in tutti gli oggetti creati nei repository. Lo strumento di reportistica dei metadati, consente di: descrivere l'origine di un dato e il processo tramite il quale esso viene derivato; individuare le dipendenze di oggetti specifici nel repository e il loro impatto su altri oggetti di Data Services (si veda Figura 6.5).

Auto Documentation Lo strumento di auto documentazione fornisce un modo veloce e completo per creare documentazione cartacea di tutti gli oggetti creati su Data Services. Questi report permettono di acquisire informazioni fondamentali per comprendere in maniera immediata l'intero processo ETL. Dopo aver creato un progetto è possibile utilizzare la sezione di documentazione automatica nella Management Console, per creare un file Word o PDF che descriva: Job, Workflow e Dataflow in versione tabellare e grafica.

Operational Dashboard La sezione Operational Dashboard fornisce una rappresentazione sulle statistiche di esecuzione dei Job. Attraverso questo strumento è possibile avere una visione sullo stato e sulle prestazioni dei Job eseguiti nei diversi repository in un dato periodo di tempo. È possibile utilizzare queste informazioni per ottimizzare e monitorare le schedulazioni dei Job, al fine di massimizzare l'efficienza e le prestazioni complessive.

Data Validation La sezione Data Validation utilizza una dashboard che permette di valutare l'affidabilità dei risultati basate su regole di valutazione create nei Job di Data Services. Attraverso le dashboard presenti in questa sezione, gli utenti del business possono rivedere velocemente le trasformazioni ed identificare potenziali inconsistenze o errori.

6.3 QlikView 11

QlikView o QV è uno strumento di reportistica per la Business Discovery prodotta dall'organizzazione QlikTech Internationl. QV permette l'analisi di dati a differenti livelli di aggregazione, tramite la creazione, la modifica e la pubblicazione di report, cruscotti direzionali o altri strumenti utili per l'analisi dei dati *user-driven*, tipica della Business Discovery. QV adotta una logica di tipo associativa, in quanto i join tra le tabelle sono effettuati sulla base della nomenclatura dei campi delle colonne.

Analisi in-memory Il sistema di reportistica QlikView utilizza una tecnica di analisi *in-memory*, in altre parole carica tutti i dati in memoria ed esegue le aggregazioni ed i join in RAM, portando ad un possibile rallentamento delle altre applicazioni che stanno girando sulla macchina. L'utilizzo della RAM non è costante, ma subisce un incremento durante l'importazione dei dati ed il caricamento di questi nei report realizzati. Una volta che i dati sono caricati, la loro esplorazione richiede una quantità di memoria costante a meno di piccole variazioni dovute alle operazioni di filtraggio. Lo svantaggio della RAM occupata viene compensato da una maggiore velocità nella disponibilità dei dati e nella loro analisi.

Flessibilità Una delle principali problematiche associate all'elaborazione OLAP tradizionale consiste nel fatto che la modifica di un'analisi comporta la modifica del cubo, operazione che può richiedere tempo. Con QV i responsabili aziendali possono visualizzare un'analisi creando o modificando dimensioni e misure in tempi ridotti. Le interfacce standard, inclusi ODBC e Web Services, consentono a QV di analizzare informazioni provenienti da qualsiasi origine e di estrarle direttamente dalla loro fonte, che può essere rappresentata da: file di testo (csv, file Excel, XML ecc..), tabelle, nonché data warehouse e data mart.

6.3.1 Architettura

Di seguito si descrivono brevemente le componenti e le funzionalità principali dello strumento.

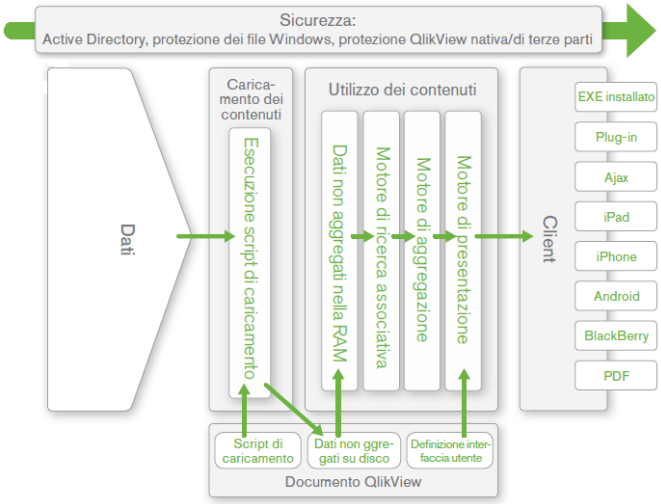


Figura 6.6: Prospettiva funzionale di QlikView

QlikView Developer Consente di definire gli script di caricamento e di visualizzazione per le applicazioni QV. Gli script di caricamento definiscono sia le connessioni alle sorgenti, sia i dati che verranno estratti. Dopo aver trasformato tali dati, sviluppatori e progettisti possono utilizzare il Developer per definire le visualizzazioni che verranno utilizzate dagli utenti finali. Gli script di caricamento sono scritti in AQL (Associative Query Language) linguaggio proprietario della QlikTech.

QlikView Server Consente di ricaricare, proteggere, gestire e distribuire agli utenti finali i contenuti dei documento QV attraverso Internet o Intranet. Questa piattaforma è integrata in QV per fornire agli utenti finali una suite di tecnologie di analisi dei dati tra loro compatibile. Il componente server di QV rappresenta il nucleo di questa tecnologia, in quanto fornisce un insieme di documenti QV gestiti a livello centrale.

Caratteristiche supportate:

- gestione utenti;
- repository per documenti;
- reload dei dati;
- autenticazione degli utenti;
- autorizzazione per l'accesso ai dati.

QlikView Publisher È un componente opzionale progettato per gestire scenari complessi di implementazione dei contenuti tipiche delle aziende

di grandi dimensioni. Il Publisher: estende ed ottimizza le funzionalità di pianificazione del Server, assicurando al contempo una maggiore protezione dei contenuti; consente di distribuire i dati salvati nei documenti QlikView agli utenti che si trovano all'interno e all'esterno dell'organizzazione. Grazie alla personalizzazione dei contenuti, ogni utente potrà visualizzare soltanto le informazioni di cui ha bisogno.

QlikView AccessPoint Con AccessPoint gli utenti finali hanno la possibilità di accedere a tutti i contenuti di QV per cui sono autorizzati. AccessPoint fornisce inoltre una serie di servizi in back-end, tra cui sessioni di bilanciamento del carico su più Server.

QlikView Client Una volta che i contenuti sono stati implementati tramite QlikView Server, sono pronti per essere utilizzati dagli utenti finali. Per QlikTech è fondamentale che i contenuti di QV vengano messi a disposizione ovunque e in qualsiasi momento. Affinché questo sia possibile QV supporta numerose tecnologie e modalità per il collegamento al Server:

- accesso al browser per poter visualizzare i report da qualsiasi PC o desktop;
- supporto di dispositivi mobili;
- accesso offline grazie al client installato;
- invio dei report tramite posta elettronica.

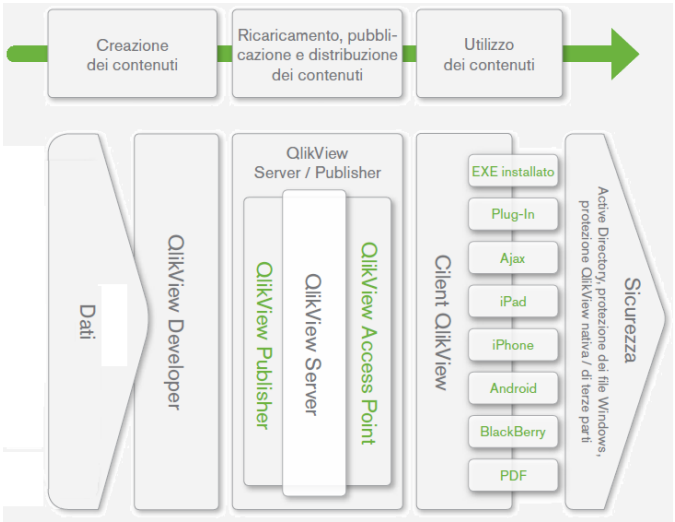


Figura 6.7: Architettura QlikView

6.3.2 Documento QlikView

Il formato file nativo di QV, chiamato anche documento QlikView, è esaustivo, portatile e facile da utilizzare. I documenti di *QlikView* includono:

- Script di caricamento. Gli script di caricamento di QV estraggono le informazioni ed elaborano i dati non aggregati che sono stati restituiti dal processo di caricamento.
- Definizioni dell'interfaccia utente per le visualizzazioni, ovvero definizioni di tabelle, grafici, cruscotti e altri elementi di reporting.
- Dati da analizzare. I documenti di QV possono anche contenere i dati da analizzare in formato compresso. In questo modo è possibile eseguire analisi offline caricando i dati in memoria a seconda della necessità.

Pubblicazione Quando un documento QV viene pubblicato sul QlikView Server, il suo contenuto viene messo a disposizione di tutti gli utenti finali che dispongono dei privilegi per accedervi. Alla prima apertura di un documento QV i dati vengono caricati in memoria. Il dataset compresso e non aggregato viene scaricato dal disco e caricato nella RAM del Server. Questo repository in-memory funge da dataset di base per il primo utente e per tutti gli altri utenti che richiedono lo stesso documento. Se per un determinato periodo di tempo non si registrano attività degli utenti, il repository viene rimosso dalla memoria.

Esplorazione Gli utenti esplorano i dati tramite selezioni. Il concetto alla base di QV è lo stato di selezione definito dall'utente. Cliccando in un documento QV, gli utenti indicano quali sono i sottoinsiemi di dati da analizzare e quali da ignorare. QV trae vantaggio dalla natura altamente indicizzata dei dataset non aggregati e presenta dinamicamente un sottoinsieme di tutti i dati disponibili per il documento a seconda dello stato di selezione, il tutto in tempo reale.

Capitolo 7

Procedure ETL

Una volta terminate le fasi di analisi dei requisiti e di progettazione del data warehouse, si procede alla creazione della base di dati di supporto alle decisioni. In questo capitolo si descrive l'implementazione delle fasi di estrazione, trasformazione e caricamento dei data mart che compongono il data warehouse finale; si presentano alcune delle problematiche riscontrate durante lo sviluppo dei *Job* Data Services e il modo in cui, tali problematiche sono state affrontate. Il capitolo si conclude con la descrizione sulle modalità di aggiornamento del data warehouse e un breve cenno all'attività di *Test* e *Tuning* affrontata durante il progetto.

7.1 Il processo ETL

Come riportato in [Albano 13], si definisce una chiara e sintetica visione del processo di ETL.

Definition 5 (ETL) *Una serie di operazioni per ottenere dati da fonti operative (fase di estrazione), pulire e preparare tali dati (fase di trasformazione) per il loro effettivo caricamento nel data warehouse (fase di caricamento).*

In generale, prima che i dati siano inseriti nel data warehouse sono sottoposti a tre operazioni:

Estrazione. Estrazione dei dati dalle applicazioni di produzione e dalle basi di dati di produzione (ERP, CRM, RDBMS, file ecc.).

Trasformazione. Fase di consolidamento dei dati: l'attività principale è di uniformare i dati, renderli coerenti con le regole imposte dal business e integrarli in un unica base di dati. Di seguito sono elencate alcune operazioni di trasformazione:

- codifica;
- aggiunta di campi e misure derivate;
- gestione delle chiavi surrogate;
- manipolazione delle stringhe;
- manipolazione del formato dei dati;
- aggregazioni;
- gestione degli errori.

Caricamento. Fase di caricamento dei dati all'interno delle tabelle che costituiscono il data mart o il data warehouse.

Nel progetto presentato in questa tesi, la fase di estrazione, trasformazione e caricamento, è realizzata mediante il software *Designer* di *SAP BODS XI 3.2* presentato nel Capitolo 6. Prima di procedere alla descrizione dell'implementazione di parte della soluzione, si fornisce un' introduzione allo strumento *Designer*, software di sviluppo di SAP Business Objects Data Services.

7.2 SAP BODS Designer

Come già riportato nel Capitolo 6, il software Designer consente di creare *Workflow*, *Dataflow* e tutte le restanti entità o oggetti (per la definizione di oggetto si veda il Paragrafo 6.2.2).

7.2.1 Job

Un Job è l'unico oggetto in Data Services che può essere eseguito e testato in modalità manuale o pianificata. Ogni Job è a sua volta composto da diversi oggetti, quali:

- **Dataflow.** Rappresenta un flusso di dati. Quest'ultimi, generalmente, partono da una tabella di origine, vengono trasformati da un oggetto *transform* e vengono caricati in una tabella di destinazione.
- **Workflows.** Contiene al suo interno oggetti che gestiscono la corretta esecuzione dei dataflow (oggetti script, conditional, Try/Catch, While ecc.).

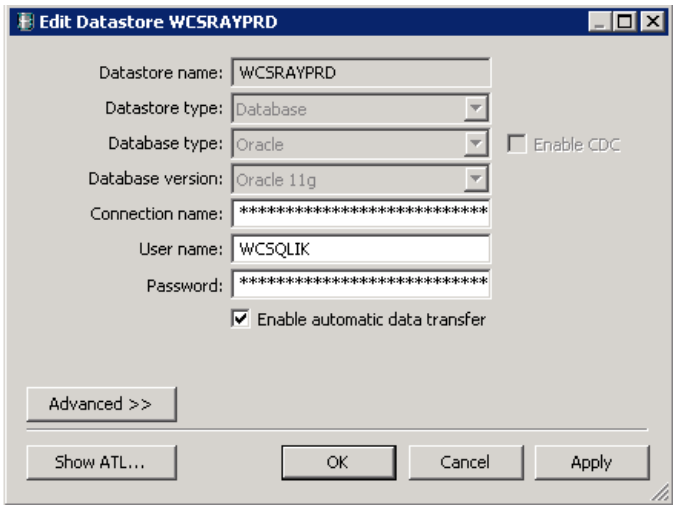


Figura 7.1: Configurazione del sistema sorgente IBM WCS

7.2.2 Datastore

L'oggetto datastore, consente di configurare le connessioni a vari tipi di sistemi di sorgente dati, tra i quali:

- base di dati;
- file di sistema;
- sistemi gestionali (J.D. Edwards, Oracle Applications, SAP Applications ecc.).

Per ogni datastore si possono definire più configurazioni alla stessa connessione. Questo permette di configurare diversi ambienti (testing, quality ecc.) in un unico oggetto datastore, per agevolare le attività di sviluppo multi-utente e quelle di integrazione dei Job nell'ambiente di produzione. In Figura 7.1 è riportata la configurazione per la connessione al database di Oracle contenente le tabelle di IBM WCS (si veda Capitolo 5).

7.2.3 File Format

Nel Designer, i file formattati possono essere utilizzati come sorgente o destinazione dei dati. L'Objects Library (si veda Paragrafo 6.2.2) memorizza il *template* del file formattato, definito dall'utente, in un oggetto utilizzabile all'interno dei dataflow. Un oggetto File Format è configurabile per alcune tipologie di file, tra cui:

- file con dati delimitati da separatori (carattere, tabulazione ecc.);
- file di testo non strutturati;

- file binari non strutturati.

In Figura 7.2 è riportata la configurazione di un oggetto File Format per l'importazione dei dati di Omniture Syte Catalyst, presenti in una cartella del file system. Mentre il nome del file è statico (*RBVISID.csv*), il suo percorso dipende dal valore impostato nella variabile globale *$G_PATH_OMNITURE*.

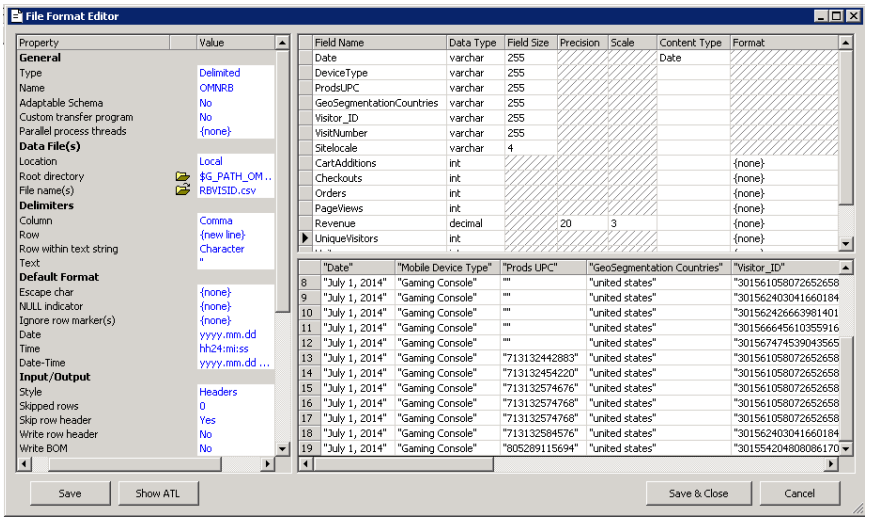


Figura 7.2: Configurazione del File Format

7.2.4 Dataflow

Un dataflow è lo spazio riservato al trasferimento di dati da una sorgente verso una destinazione applicando delle trasformazioni. Un dataflow può essere posizionato all'interno dell'oggetto workflow. Dal workflow, un dataflow può inviare e ricevere informazioni di altri oggetti, attraverso parametri definiti dallo sviluppatore.

Figura 7.3: Dataflow di SAP BODS

Supponendo di voler popolare una tabella con dei nuovi dati, provenienti da due tabelle, il dataflow sarà definito come in Figura 7.4:

- due tabelle sorgenti;

- un operatore di join definito in un oggetto *transform*;
- una tabella di destinazione, dove saranno inserite le nuove righe.

Il verso dei collegamenti degli oggetti, all'interno del dataflow, impone l'ordine del flusso dati.

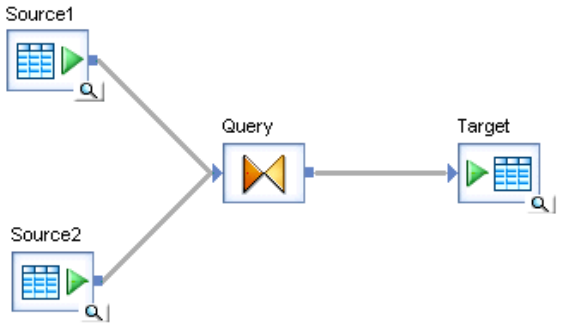


Figura 7.4: Contenuto di un dataflows

7.2.5 Workflow

Un workflow definisce il processo decisionale per l'esecuzione dei dataflow. In generale, durante l'esecuzione di un Job, il ruolo di un workflow è di organizzare l'esecuzione temporale dei dataflow, gestire la configurazione dell'ambiente di esecuzione prima e dopo che i dataflow siano terminati e prendere percorsi alternativi in presenza di determinate condizioni.

Figura 7.5: Workflow di SAP BODS

Come si nota dalla Fiugra 7.6, all'interno di un workflow si possono definire script e catturare errori che non permettono la corretta esecuzione dei dataflow o altre situazioni che influenzano il normale flusso dei dati.

Figura 7.6: Esempio di un workflow

7.2.6 Transform

Un oggetto di tipo transform riceve un insieme di dati, in formato tabellare, da un oggetto sorgente (tabella di una base di dati, file di testo, file xml ecc..) effettua delle operazioni e produce un insieme di dati in output. Gli oggetti transform sono raggruppati nelle seguenti categorie:

- **Data Integrator:** permettono di estrarre, trasformare e caricare dati. Gli oggetti Data Integrator supportano le attività di aggregazione dei dati, del caricamento e aggiornamento del data warehouse;
- **Data Quality:** aiutano a migliorare la qualità dei dati. Gli oggetti Data Quality supportano le attività di standardizzazione, correzione, arricchimento e consolidamento dei dati;
- **Platform:** oggetti necessari per le operazioni di movimentazione dati più comuni. In generale, gli oggetti Platform generano operazioni tradotte in query SQL;
- **Text Data Processing:** aiutano ad estrarre specifiche informazioni da file di testo. Gli oggetti Data Processing permettono di identificare ed estrarre entità e fatti rilevanti per l'utente finale.

7.2.7 Variabili e parametri

L'utilizzo di variabili locali e globali, permette di incrementare la flessibilità e la riusabilità degli oggetti worfkflow e dataflow. Le variabili, utilizzate all'interno dei dataflow, facilitano la manipolazione dei dati. Le variabili possono essere utilizzate, ad esempio, in una funzione o in una clausola *where* di un oggetto *query transform.*

Una variabile locale ha una visibilità ristretta all'oggetto nel quale è stato creato. Una variabile globale, invece, ha una visibilità limitata al Job nel quale è stata creata e contrariamente a quanto accade per le variabili locali, non richiede l'utilizzo di parametri. La valorizzazione di una variabile globale o locale avviene all'interno degli oggetti di tipo script.

La Tabella 7.1 elenca il livello di definizione e di visibilità delle variabili locali, globali e dei parametri.

7.2.8 Script

Uno script è un oggetto non riusabile (si veda Paragrafo 6.2.2), definito in un workflow, per chiamare una funzione o assegnare un valore ad una variabile. In sintesi, uno script può contenere i seguenti *statements*:

- chiamate di funzioni;

Livello di definizione	**Variabili e parametri**	**Livello di visibilità**
Job	Variabili globali	Qualsiasi oggetto nel Job
Workflow	Variabili locali	Nel workflow stesso o un workflow/dataflow definito al suo interno
Workflow	Parametri	Oggetto padre del workflow
Dataflow	Parametri	Clusola *where*, column mapping o funzione interna al dataflow

Tabella 7.1: Tabella riepilogativa delle dimensioni

- dichiarazioni *If-then-else*;
- dichiarazioni *While-do*;
- valorizzazione di variabili.

7.3 Metodologia di sviluppo

Durante l'incontro con i responsabili IT dell'azienda committente, è stato rivisitato l'ambiente di sviluppo Data Services messo a disposizione. La problematica emersa è stata la mancanza di un ambiente di sviluppo, in grado di eseguire i Job nei Server di produzione (si veda Paragrafo 6.2.2) prima del loro rilascio nel local repository di produzione. Il collaudo dei Job, è un procedimento necessario per individuare le carenze in termini di correttezza e affidabilità delle operazioni ETL senza compromettere l'ambiente di produzione.

Al fine di risolvere la problematica riscontrata, è stato proposto e realizzato, un repository centrale di sviluppo che, oltre a soddisfare le caratteristiche di ambiente di sviluppo, agevola la gestione del versionamento del team di lavoro, centralizzando i Job e mantenendo il *versioning* delle modifiche effettuate nel tempo. Si elencano, di seguito, gli elementi creati che compongono il nuovo ambiente (si veda Figura 7.7):

- Un local repository per ogni sviluppatore (BODI_DEV). I Job sono eseguibili sui Server di sviluppo;
- Un central repository di sviluppo (BODI_DEV_CENTRAL) per la gestione del versionamento dei Job nelle attività di sviluppo quotidiano.
- Un local repository di pre-produzione (BODI_PREPROD) che rappresenta l'ambiente di collaudo. I Job all'interno del repository sono eseguibili sui Server di produzione.

Alla luce della nuova architettura di sviluppo, è stato condiviso e convalidato, il nuovo processo di rilascio in produzione dei Job Data Services:

- Operazione di *commit* (o check-in): si copiano le modifiche fatte sui Job in locale nel central repository di sviluppo (bodi_dev_central);
- Operazione di *update* (o get latest version): si copiano le modifiche salvate nel central repository di sviluppo nel local repository di preproduzione (bodi_preprod);
- Una volta che i Job sono idonei per il rilascio nell'ambiente di produzione si effettua un'operazione di commit nel central repository ufficiale (bodi_central);
- Dal local repository di produzione (bodi_production) i Job sono schedulati per la loro esecuzione periodica.

Attenersi a questo processo di rilascio, permette di garantire un'esecuzione corretta dei Job in ambiente di produzione, dato che l'identificazione di incorrettezze viene anticipata ad una fase intermedia.

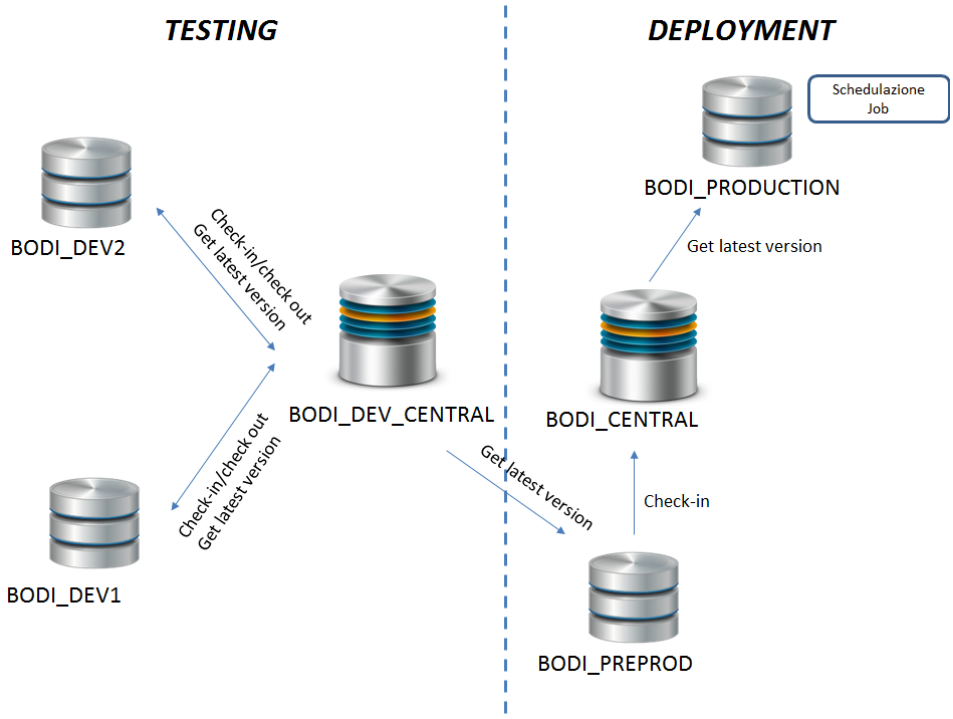


Figura 7.7: Metodologia di sviluppo in SAP Business Objects Data Services

7.4 Fasi di sviluppo

Si elencano, in prima analisi, le macro fasi che costituiscono il processo di progettazione e sviluppo dell'ETL:

1. data profiling;
2. progettazione della base di dati;
3. progettazione e sviluppo dei Job in Data Services;
4. test e tuning.

Con riferimento al progetto presentato in questa tesi, vederemo, nelle seguenti sezioni, come sono state affrontate le varie fasi di sviluppo e il tipo di soluzioni adottate.

7.5 Data Profiling

È stata eseguita una prima importazione di tutte le tabelle, presenti nei sistemi di origine, all'interno del RDBMS di Netezza, allo scopo di verificare:

- il livello di pulizia del dato;
- la correttezza delle chiavi primarie all'interno delle tabelle e la loro corretta valorizzazione su tutti i record;
- l'integrità referenziale rispettata fra le entità evento e le entità componente (si veda la Sezione 5.2);
- attributi di descrizione presenti e non ripetuti;
- corrispondenza fra le relazioni presenti nei modelli ideati in fase di progettazione dei data mart e le relazioni presenti nei sistemi di partenza. Le relazioni *1:n* presenti nei modelli devono esistere anche all'interno dei sistemi di origine e non essere violate, altrimenti o il modello non rispecchia le relazioni effettive e deve quindi essere rivisitato, o la sorgente informativa consente, al suo interno, forzature anomale che andranno gestite all'interno del data warehouse.

7.6 Progettazione delle basi di dati

Il sistema di basi di dati che contiene il data warehouse è suddiviso in:

- tabelle temporanee;
- tabelle delle anomalie;

- tabelle di staging area;
- tabelle trattate manualmente dagli utenti;
- star schema.

Il sistema di basi di dati dedicato alla gestione dei Job ETL è suddiviso in:

- tabella di esecuzione;
- tabella delle procedure;
- tabelle dello storico.

Le tabelle temporanee ospitano i dati provenienti dalle sorgenti informative. Ne esiste una per ogni tabella o file del sistema di origine che deve essere importata. Le tabelle temporanee contengono chiavi e relazioni logiche ereditate dal sistema di origine. Una volta che i dati sono stati trasferiti nella staging area, i record contenuti nelle tabelle temporanee vengono eliminati.

Le tabelle di staging area, contengono i dati provenienti dalle tabelle temporanee. Le chiavi delle tabelle di staging possono essere surrogate nel caso in cui la chiave originale occupi troppo spazio e renda critiche le dimensioni della tabella dei fatti. In generale, le tabelle di staging area sono utilizzate per la pulizia e la certificazione dei dati di origine.

Le tabelle manuali contengono i dati gestiti manualmente dagli utenti del business, quindi informazioni non presenti nei sistemi operazionali.

Lo star schema è l'area del data warehouse interrogata direttamente dal sistema di reporting. I dati sono rappresentati nella forma congeniale all'interrogazione del sistema di reportistica. Le tabelle dello star schema sono organizzate in schemi a stella. Ogni modello a stella contiene una tabella dei fatti e diverse tabelle dimensionali (si veda Figura 5.10). Per ogni dimensione esiste un unica tabella di lookup, dato che la presenza dell'informazione denormalizzata rende più rapida la ricerca e la presentazione dei dati.

Le tabelle per la gestione dei Job saranno descritte nel seguito del capitolo.

7.7 Job Data Services

Ogni Job è caratterizzato da un Workflow che contiene un insieme di *procedure*. Le procedure sono oggetti di tipo Conditional che gestiscono l'esecuzione e la configurazione di un insieme di dataflow correlati fra di loro. In generale, per ogni data mart progettato (si veda Capitolo 4), sono stati creati

tre Job che rispecchiano le fasi di estrazione, trasformazione e caricamento. I Job hanno una struttura standard, ovvero contengono un workflow che inizializza, storicizza e coordina la corretta esecuzione delle procedure (si veda Figura 7.8).

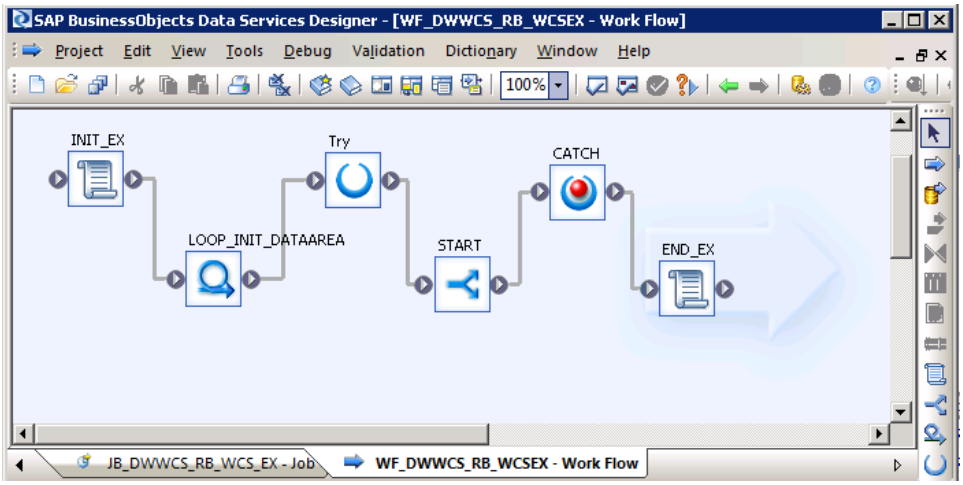


Figura 7.8: Struttura principale di un Workflow

Così come i Job, anche le procedure, contenute nell'oggetto *start* (si veda Figura 7.9), presentano una struttura standard che consiste nelle seguenti operazioni:

- lettura dalla tabella di registro per l'autorizzazione all'esecuzione dei dataflow;
- esecuzione dei dataflow;
- chiusura della procedura ed aggiornamento della tabella di registro;
- storicizzazione della procedura.

Vediamo di seguito, alcune delle operazioni create con lo strumento Designer, che prendono parte all'implementazione del data warehouse finale.

7.7.1 Estrazione

Il Job di estrazione contiene le seguenti procedure:

- **Import:** importazione delle tabelle e dei file dei sistemi di origine;
- **Export:** trasferimento dei dati dalle tabelle temporanee alla staging area.

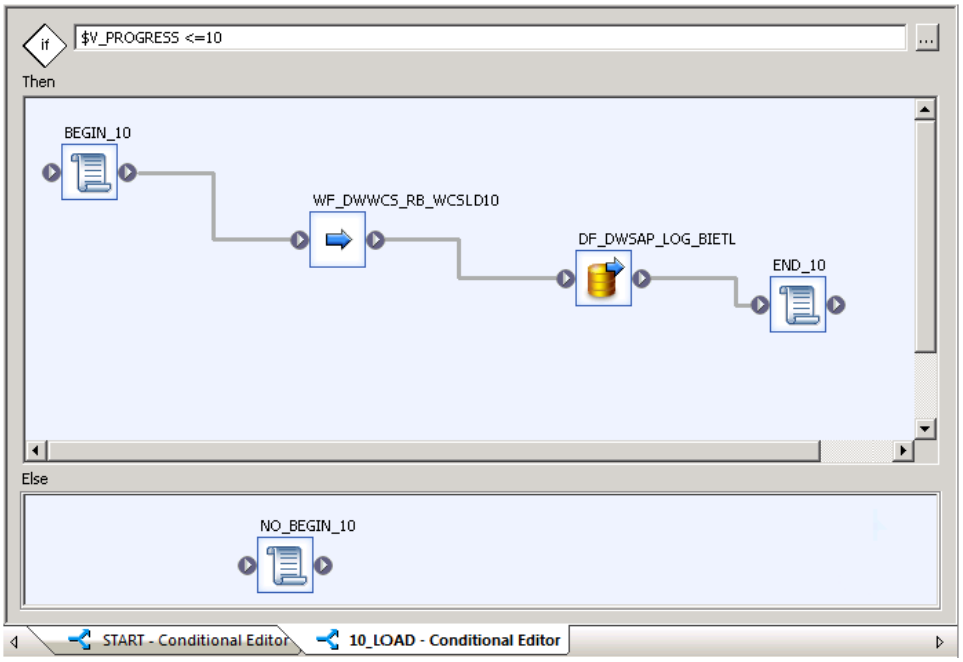


Figura 7.9: Struttura di una procedura

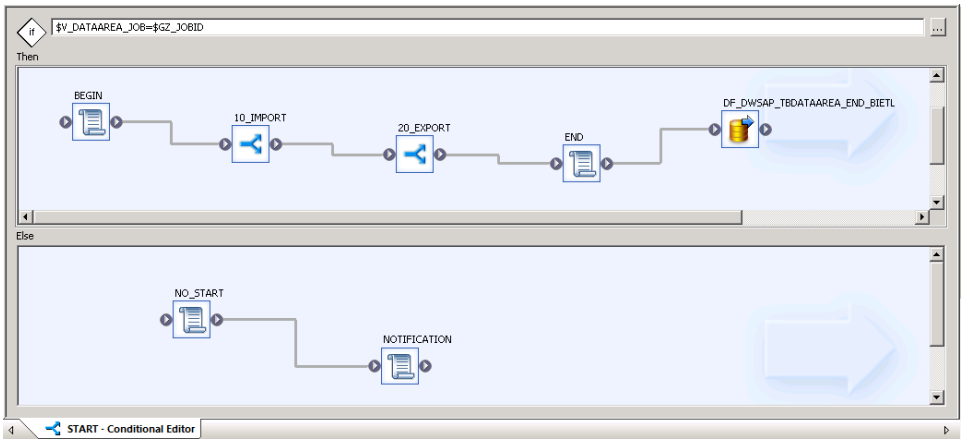


Figura 7.10: Job di estrazione

Import

Le operazioni di importazione di ogni tabella o file, sono contenute in questa procedura. In Figura 7.11 è riportato un esempio della procedura *Import* riferita al Job di estrazione di Web Commerce Suite. L'oggetto condizionale *Period* effettua: un'importazione giornaliera o massiva dei dati dell'entità

evento, i quali andranno a popolare la tabella finale dei fatti; un'importazione massiva dei dati dell'entità componente, i quali andranno a popolare le tabelle finali delle anagrafiche. Nel caso in cui ci sia un errore di connessione al database Oracle di WCS o un'assenza del file di Omniture Site Catalyst nella cartella del file system, gli oggetti di tipo *Try/Catch*: catturano e segnalano, tramite mail, il tipo di errore generato; aggiornano la tabella di registro delle procedure; terminano l'esecuzione del Job.

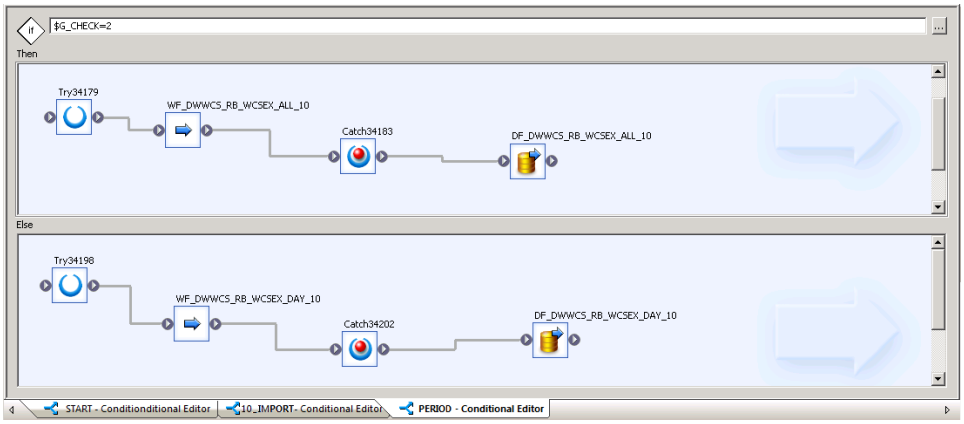


Figura 7.11: Importazione delle tabelle di IBM Web Commerce Suite

In Figura 7.12 è riportato il contenuto del dataflow di importazione dei dati riguardanti Omniture Site Catalyst. L'oggetto File Format è descritto nel Paragrafo 7.2.3. L'oggetto *query transform* assicura che non vengano inserite nelle tabelle temporanee caratteri anomali che potrebbero generare errori non gestibili in fase di importazione.

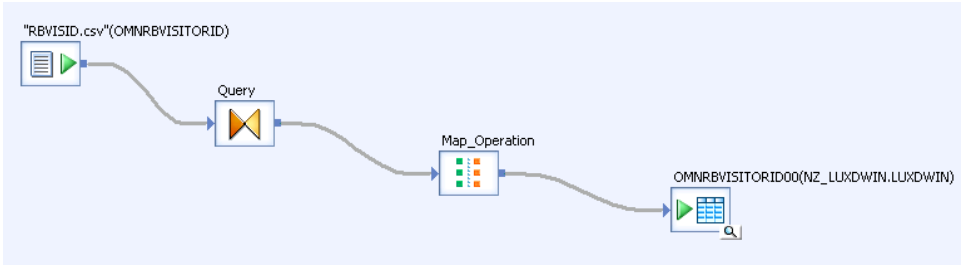


Figura 7.12: Estrazione del file di Omniture Site Catalyst

Export

All'interno della procedura Export sono contenute le operazioni di trasferimento dei file, dalle tabelle temporanee a quelle di staging area. Come si nota in Figura 7.13, ogni tabella temporanea è associata ad un data-flow che attua un semplice caricamento dei dati senza alcuna operazione di trasformazione.

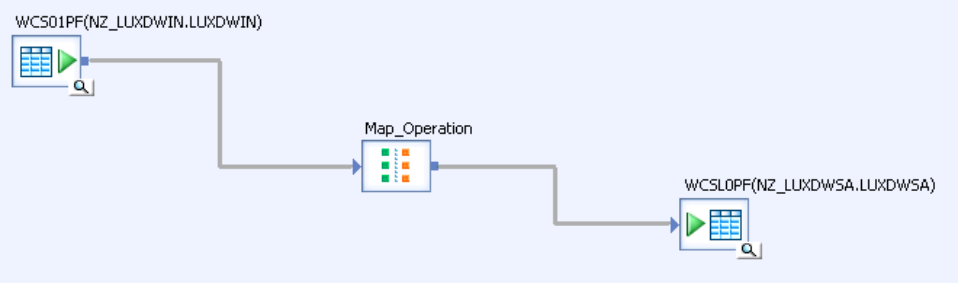


Figura 7.13: Trasferimento in staging area della tabella degli ordini di WCS

7.7.2 Staging Area

La staging area è definita come un'area temporanea di memorizzazione, che si colloca fra la sorgente dati e il data warehouse. All'interno di quest'area avvengono le tipiche operazioni di trasformazione, propedeutiche al caricamento del data warehouse. Alcuni dei vantaggi derivanti dall'utilizzo della staging area sono riassumibili nei seguenti punti:

- Possibilità di utilizzare tabelle ausiliarie, non utili per le analisi, ma necessarie per completare determinate operazioni o arricchire le informazioni di cui si dispone.
- Possibilità di effettuare trasformazioni in diversi passi, in modo da ottenere una sequenza strutturata, chiara e leggibile a discapito di un maggior tempo di elaborazione.
- Possibilità di avere, ove necessario, una prima fase di pre-aggregazione dei dati.

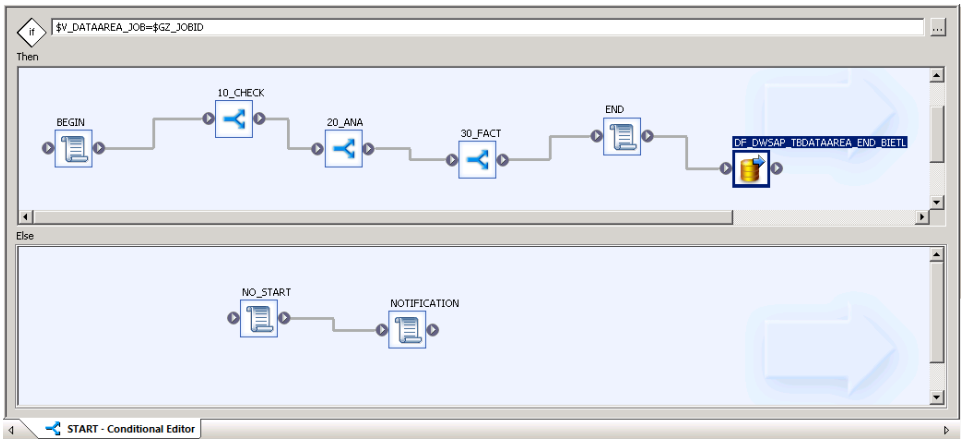


Figura 7.14: Job di trasformazione

Le operazioni tipiche di staging area sono presenti nelle varie procedure del Job di trasformazione (Figura 7.14). In particolare, il Job è stato suddiviso in tre procedure:

- **Check:** procedura di controllo e gestione dei vincoli d'integrità;
- **Ana:** popolamento delle tabelle finali di anagrafica di staging area;
- **Fact:** popolamento delle tabelle finali dei fatti di staging area.

Check

Si elencano di seguito i tipici vincoli di integrità, ovvero prorietà che devono essere soddisfatte da tutti i record presenti nelle tabelle che compongono il datawarehouse:

- **Violazione di chiave primaria.** La violazione di chiave primaria si verifica quando il campo chiave non è valorizzato, oppure esistono due record con lo stesso valore nel campo chiave. I record che non hanno un valore nel campo di chiave primaria non possono essere inseriti nel data warehouse e di conseguenza sono scartati.
- **Violazione di integrità referenziale.** La violazione d'integrità referenziale si verifica quando, un valore nella tabella dei fatti dichiarato come chiave esterna, contiene un valore corrispondente a nessun valore presente nella chiave primaria della tabella delle dimensioni.
- **Mancata valorizzazione degli attributi descrittivi.** I record che non hanno descrizione valorizzata sono segnalati e inseriti all'interno

del data warehouse. La descrizione mancante è sostituita con una descrizione fittizia.

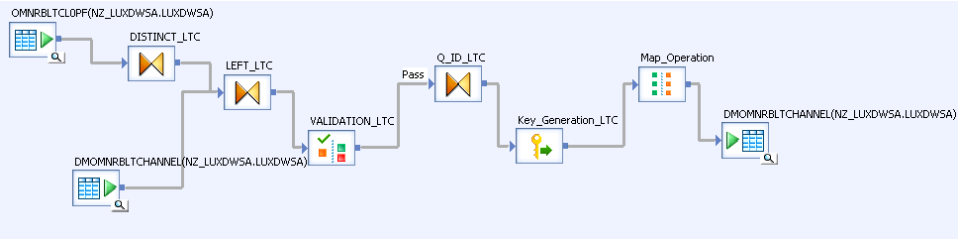


Figura 7.15: Popolamento della tabella di anagrafica in staging area

Per ovviare a queste anomalie, il popolamento di una tabella di anagrafica va sempre eseguito facendo un raggruppamento sulla chiave. In questo modo si ha la certezza di non violare la chiave primaria nelle tabelle di staging area.

In Figura 7.15 è presentato un esempio di aggiornamento della tabella finale di anagrafica di staging area *DMOMNRBLTCHANNEL*.

La tabella *OMNRBLTCL0PF* contiene i dati aggiornati relativi all'attributo dimensionale *Last touch channel*. L'operatore *Transform* confronta *OMNRBLTCL0PF* con la tabella finale di anagrafica di staging area e restituisce in output i valori non presenti in quest'ultima tabella. Infine, l'operatore *Key Generator* associa una chiave surrogata ad ogni valore ricevuto in input e restituisce in output le nuove righe da inserire nella tabella finale di anagrafica di staging area.

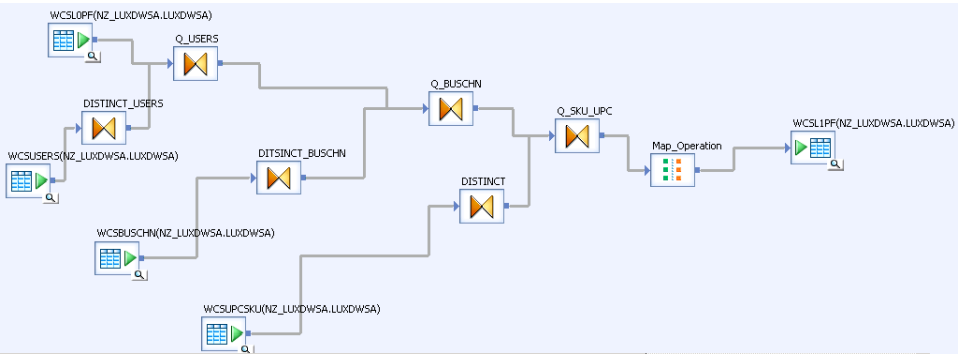


Figura 7.16: Controllo di integrità referenziale (WCS)

Per il controllo sulla violazione di integrità referenziale (si veda Figura 7.16), è necessario incrociare i campi di chiave esterna dell'entità componente, con gli stessi campi che sono chiave nell'entità evento. Le relazioni corrette sono mantenute, mentre le chiavi esterne mancanti vanno sostituite con elementi fittizi. Un attributo fittizio, in anagrafica, permette di costruire le gerarchie mancanti all'interno delle dimensioni. In questo modo gli elementi dei fatti che violano l'integrità referenziale, sono comunque caricati, associandoli il valore fittizio.

Popolamento delle tabelle finali di staging area

Il popolamento della tabelle finali di anagrafica di staging area avviene con i soli record che hanno superato i controlli imposti dai vincoli di cui sopra. Il popolamento della tabella finale di staging area avviene successivamente all'aggiornamento delle anagrafiche. Come già precedentemente illustrato, i riferimenti a chiavi esterne mancanti vengono sostituiti con riferimenti ad elementi fittizi creati *ad hoc* nelle anagrafiche. Le tabelle dei fatti si incrociano con le tabelle di anagrafica solamente tramite left join (si veda Figura 7.17).

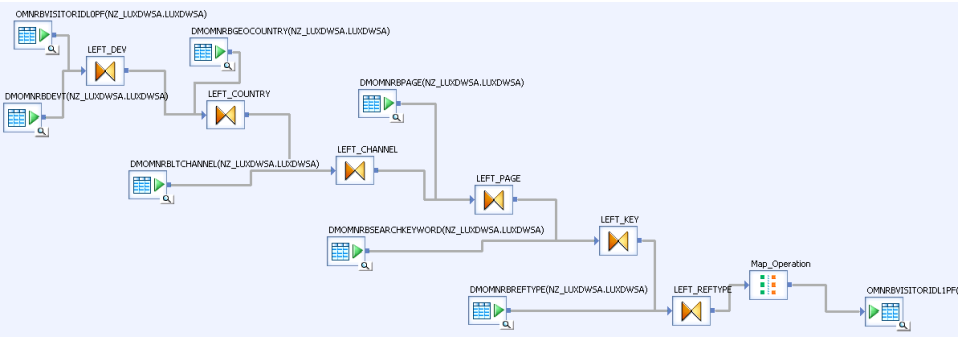


Figura 7.17: Popolamento della tabella finale dei fatti di staging area

7.7.3 Popolamento del data warehouse

La procedura *Load* presente nel Job di caricamento non necessita di controlli, dato che interroga una base dati già certificata: la staging area. Il popolamento del data warehouse avviene in *full refresh* a partire dalle tabelle finali di staging area.

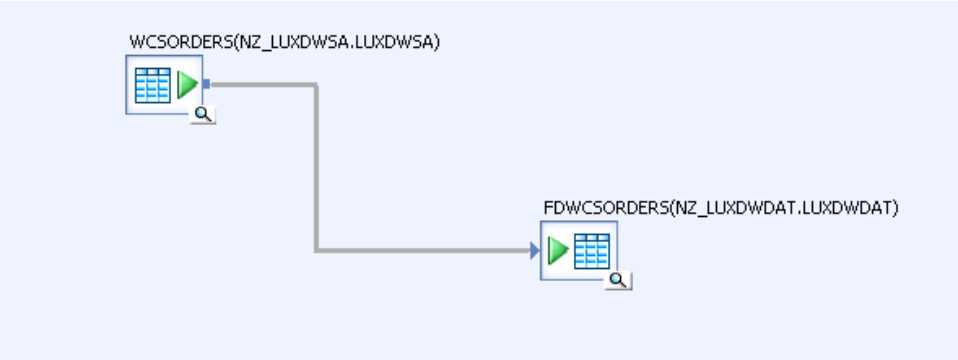
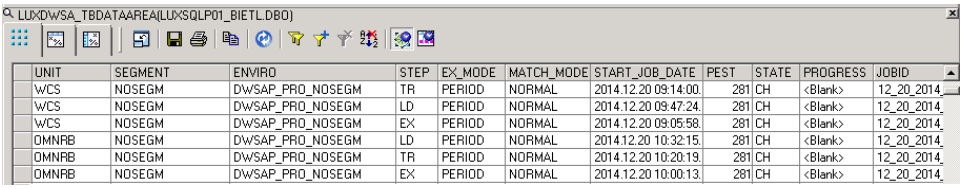


Figura 7.18: Popolamento della tabella dei fatti del data warehouse

7.8 Gestione delle procedure

In questa sezione si descrivono le logiche di controllo e di esecuzione delle procedure. Lo scopo di tale implementazione è di avere un monitoraggio dettagliato sullo stato di esecuzione delle procedure e, in caso di errore, continuare o meno l'esecuzione del Job a seconda della reversibilità dell'errore riscontrato.

Tabella di esecuzione


LUXDWSA_TBDATAAREA(LUXSQLP01_BIETL.DBO)

UNIT	SEGMENT	ENVIRO	STEP	EX_MODE	MATCH_MODE	START_JOB_DATE	PEST	STATE	PROGRESS	JOBID
WCS	NOSEGM	DWSAP_PRO_NOSEGM	TR	PERIOD	NORMAL	2014.12.20 09:14:00.	281	CH	<Blank>	12_20_2014_
WCS	NOSEGM	DWSAP_PRO_NOSEGM	LD	PERIOD	NORMAL	2014.12.20 09:47:24.	281	CH	<Blank>	12_20_2014_
WCS	NOSEGM	DWSAP_PRO_NOSEGM	EX	PERIOD	NORMAL	2014.12.20 09:05:58.	281	CH	<Blank>	12_20_2014_
OMNRB	NOSEGM	DWSAP_PRO_NOSEGM	LD	PERIOD	NORMAL	2014.12.20 10:32:15.	281	CH	<Blank>	12_20_2014_
OMNRB	NOSEGM	DWSAP_PRO_NOSEGM	TR	PERIOD	NORMAL	2014.12.20 10:20:19.	281	CH	<Blank>	12_20_2014_
OMNRB	NOSEGM	DWSAP_PRO_NOSEGM	EX	PERIOD	NORMAL	2014.12.20 10:00:13.	281	CH	<Blank>	12_20_2014_

Figura 7.19: Tabella di esecuzione delle procedure

La tabella di esecuzione (*TBDATAAREA*) contiene le informazioni del Job in stato di avanzamento, si elencano le principali:

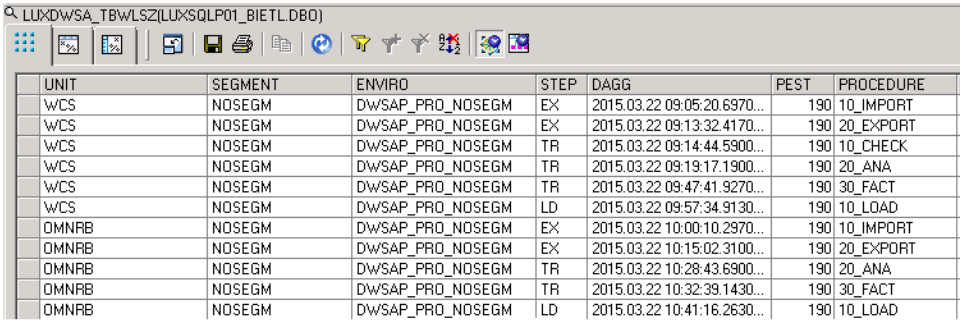
- informazioni sull'ambiente di esecuzione;
- modalità di estrazione dei dati sorgente
- job data e ora di inizio;
- stato di avanzamento;
- identificatore univoco.

La tabella di esecuzione viene aggiornata e letta ogni qual volta inizia o termina il Job.

A completamento del Job di estrazione, l'ultima riga inserita nella tabella *TBDATAAREA*, consentirà l'esecuzione del Job di trasferimento e così di seuito per il Job di caricamento.

Normalmente le operazioni di *update* e *select* su questa tabella vengono effettuate mediante l'utilizzo di funzioni richiamate all'interno di oggetti di tipo script.

Tabella delle procedure


LUXDWSA_TBWLSZ(LUXSQLP01_BIETL.DBO)

UNIT	SEGMENT	ENVIRO	STEP	DAGG	PEST	PROCEDURE
WCS	NOSEGM	DWSAP_PRO_NOSEGM	EX	2015.03.22 09:05:20.6970...	190	10_IMPORT
WCS	NOSEGM	DWSAP_PRO_NOSEGM	EX	2015.03.22 09:13:32.4170...	190	20_EXPORT
WCS	NOSEGM	DWSAP_PRO_NOSEGM	TR	2015.03.22 09:14:44.5900...	190	10_CHECK
WCS	NOSEGM	DWSAP_PRO_NOSEGM	TR	2015.03.22 09:19:17.1900...	190	20_ANA
WCS	NOSEGM	DWSAP_PRO_NOSEGM	TR	2015.03.22 09:47:41.9270...	190	30_FACT
WCS	NOSEGM	DWSAP_PRO_NOSEGM	LD	2015.03.22 09:57:34.9130...	190	10_LOAD
OMNRB	NOSEGM	DWSAP_PRO_NOSEGM	EX	2015.03.22 10:00:10.2970...	190	10_IMPORT
OMNRB	NOSEGM	DWSAP_PRO_NOSEGM	EX	2015.03.22 10:15:02.3100...	190	20_EXPORT
OMNRB	NOSEGM	DWSAP_PRO_NOSEGM	TR	2015.03.22 10:28:43.6900...	190	20_ANA
OMNRB	NOSEGM	DWSAP_PRO_NOSEGM	TR	2015.03.22 10:32:39.1430...	190	30_FACT
OMNRB	NOSEGM	DWSAP_PRO_NOSEGM	LD	2015.03.22 10:41:16.2630...	190	10_LOAD

Figura 7.20: Tabella di registro di esecuzione

La tabella delle procedure (*TBWLSZ*) contiene le informazioni della procedura in stato di avanzamento, si elencano le principali:

- informazioni sull'ambiente di esecuzione del Job;
- Job di appartenenza della procedura;
- data e ora di esecuzione della procedura;
- nome della procedura eseguita.

La tabella delle procedure viene aggiornata e letta ogni qual volta inizia o termina una procedura interna al Job. Le operazioni di *update* e *select* su questa tabella vengono effettuate mediante l'utilizzo di un dataflow condiviso da tutte le procedure.

Nel momento in cui viene aggiunta una riga nella tabella delle procedure, la stessa riga viene aggiunta nella tabella dello storico (TBLSWZT).

7.9 Test & Tuning

L'attività di testing è fondamentale per garantire qualità al lavoro svolto. I test sono mirati ad assicurare, oltre all'integrità dei dati che transitano dalla sorgente al data warehouse e all'applicazione di reportistica, anche una corretta compilazione delle anagrafiche, per i motivi spiegati nel Paragrafo 7.7.2. I test sono stati svolti durante la progettazione dei dataflow di Data Services e al termine del lavoro.

I test sui dataflow di popolamento delle anagrafiche e delle tabelle dei fatti vengono compiuti confrontando:

- numero di record della tabella di origine;
- numero di record scartati;
- numero di record dell'anagrafica di destinazione.

I record scartati sono controllati singolarmente per verificare che contengano effettivamente delle anomalie. Sulle tabelle dei fatti sono stati eseguiti, inoltre, test di confronto sul valore totale delle misure.

Al termine del lavoro, una volta che anche gli script di caricamento di Qlikview sono stati eseguiti, si controllano i totali delle misure sulle principali spaccature delle diverse dimensioni nel sistema di reportistica. Questi totali vengono confrontati con gli stessi totali calcolati sulle tabelle del RDBMS di Netezza.

7.9.1 Performance e esecuzione dei Job

Come già anticipato, ogni fase di ETL è associata ad un Job di Data Services. Per una decisione interna dei responsabili IT dell'azienda committente, i Workflow principali contenuti nei rispettivi Job di estrazione, trasformazione e caricamento (si veda Sezione 7.7) sono stati inclusi in un unico Job. Di conseguenza, il numero di Job Data Services è ridotto a due, un Job per l'ETL dei dati relativi a IBM WCS ed uno per l'ETL dei dati relativi ad Omniture Site Catalyst.

In prima istanza, viene effettuato un caricamento *full* della base di dati di staging area e del data warehouse. Per i dati provenienti da IBM WCS, i successivi aggiornamenti avvengono tramite un caricamento incrementale, il quale inserisce solo i record che hanno una data di modifica nella sorgente superiore a quella presente nel data mart. Per i dati provenienti da Omniture viene effettuato un aggiornamento che inserisce i record relativi ad un singolo giorno (si veda Paragrafo 5.1.1).

Nella Tabella seguente vengono riportati i tempi medi di esecuzione ETL dei vari tipi di caricamento:

		Tempo medio di caricamento	
JOB Data Services	**Descrizione**	**Full**	**Incremental**
JB_DWWCS_RB_WCS	ETL dei dati di IBM WCS	$\sim$ 4 ore	$\sim$ 15 minuti
JB_DWOMN_RB_OMN	ETL dei dati di Omniture Site Catalyst	$\sim$ 8 ore	$\sim$ 30 minuti

Tabella 7.2: Tempo medio di caricamento dei dati

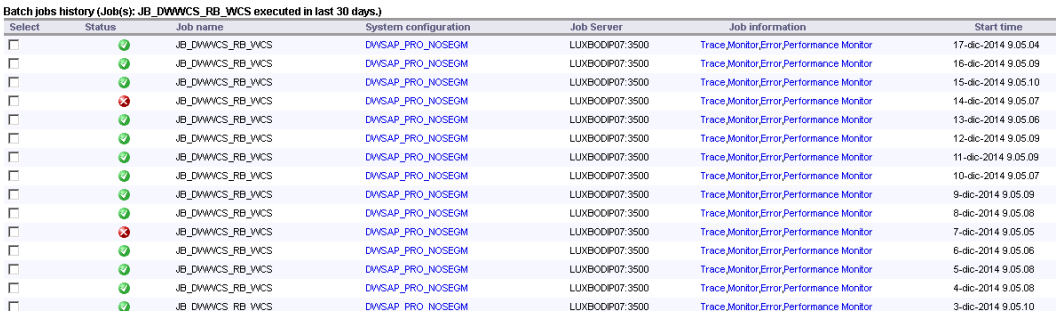
I due Job sono schedulati per essere eseguiti una volta al giorno. Il Job *JB_DWWCS_RB_WCS* alle ore 9:05 AM, momento in cui è ridotto al minimo il carico di lavoro sui server. Il Job *JB_DWOMN_RB_OMN* è schedulato alle ore 10:00 AM, momento in cui è disponibile il file fornito da Omniture Site Catalyst (si veda Paragrafo 5.1.1).

7.10 Gestione dei Job

La gestione dei Job avviene tramite il Central Management Console di SAP BO Data Services, l'interfaccia di amministrazione attraverso la quale gli utenti schedulano e monitorano i caricamenti (si veda Figura 7.21). Inoltre, lo strumento CMS è utilizzato per misurare lo stato di salute del data warehouse, infatti è possibile:

- monitorare i tempi di esecuzione dei vari Job;
- controllare eventuali *log* degli errori prodotti dall'esecuzione dei Job;
- monitorare il numero di record presenti nelle tabelle dello star schema.

In particolare, monitorare il numero di record presenti nelle tabelle più critiche serve a tenere una storia dell'attività del data warehouse per valutarne l'andamento e fare delle previsioni di crescita delle dimensioni della base di dati e dei tempi di caricamento.



Batch jobs history (Job(s): JB_DWWCS_RB_WCS executed in last 30 days.)

Select	Status	Job name	System configuration	Job Server	Job information	Start time
☐	✓	JB_DWWCS_RB_WCS	DWSAP_PRO_NOSEGM	LUXBODIP07:3500	Trace,Monitor,Error,Performance Monitor	17-dic-2014 9.05.04
☐	✓	JB_DWWCS_RB_WCS	DWSAP_PRO_NOSEGM	LUXBODIP07:3500	Trace,Monitor,Error,Performance Monitor	16-dic-2014 9.05.09
☐	✓	JB_DWWCS_RB_WCS	DWSAP_PRO_NOSEGM	LUXBODIP07:3500	Trace,Monitor,Error,Performance Monitor	15-dic-2014 9.05.10
☐	✗	JB_DWWCS_RB_WCS	DWSAP_PRO_NOSEGM	LUXBODIP07:3500	Trace,Monitor,Error,Performance Monitor	14-dic-2014 9.05.07
☐	✓	JB_DWWCS_RB_WCS	DWSAP_PRO_NOSEGM	LUXBODIP07:3500	Trace,Monitor,Error,Performance Monitor	13-dic-2014 9.05.06
☐	✓	JB_DWWCS_RB_WCS	DWSAP_PRO_NOSEGM	LUXBODIP07:3500	Trace,Monitor,Error,Performance Monitor	12-dic-2014 9.05.09
☐	✓	JB_DWWCS_RB_WCS	DWSAP_PRO_NOSEGM	LUXBODIP07:3500	Trace,Monitor,Error,Performance Monitor	11-dic-2014 9.05.09
☐	✓	JB_DWWCS_RB_WCS	DWSAP_PRO_NOSEGM	LUXBODIP07:3500	Trace,Monitor,Error,Performance Monitor	10-dic-2014 9.05.07
☐	✓	JB_DWWCS_RB_WCS	DWSAP_PRO_NOSEGM	LUXBODIP07:3500	Trace,Monitor,Error,Performance Monitor	9-dic-2014 9.05.09
☐	✓	JB_DWWCS_RB_WCS	DWSAP_PRO_NOSEGM	LUXBODIP07:3500	Trace,Monitor,Error,Performance Monitor	8-dic-2014 9.05.08
☐	✗	JB_DWWCS_RB_WCS	DWSAP_PRO_NOSEGM	LUXBODIP07:3500	Trace,Monitor,Error,Performance Monitor	7-dic-2014 9.05.05
☐	✓	JB_DWWCS_RB_WCS	DWSAP_PRO_NOSEGM	LUXBODIP07:3500	Trace,Monitor,Error,Performance Monitor	6-dic-2014 9.05.06
☐	✓	JB_DWWCS_RB_WCS	DWSAP_PRO_NOSEGM	LUXBODIP07:3500	Trace,Monitor,Error,Performance Monitor	5-dic-2014 9.05.08
☐	✓	JB_DWWCS_RB_WCS	DWSAP_PRO_NOSEGM	LUXBODIP07:3500	Trace,Monitor,Error,Performance Monitor	4-dic-2014 9.05.08
☐	✓	JB_DWWCS_RB_WCS	DWSAP_PRO_NOSEGM	LUXBODIP07:3500	Trace,Monitor,Error,Performance Monitor	3-dic-2014 9.05.10

Figura 7.21: Monitoraggio di un Job

Capitolo 8

Reportistica

Nel presente capitolo viene descritta la logica e la procedura utilizzata per la realizzazione dei report tramite lo strumento Qlikview 11 (si veda Capitolo 6). La reportistica prodotta è di natura dinamica, ovvero dashboard caratterizzate da facilità di lettura, immediatezza e possibilità di modificare i filtri predefiniti di ogni report.

Il capitolo inizia con una breve introduzione al concetto di reportistica per il supporto ai processi decisionali, segue la descrizione sulle modalità di caricamento dei dati dal data warehouse realizzato e le funzionalità introdotte nello strumento di reportistica, per facilitarne l'utilizzo. Il capitolo si conclude con la descrizione di alcuni report sviluppati.

8.1 Reporting

Per agevolare l'interpretazione dei risultati delle analisi dei dati è importante presentarli in modo opportuno. Esistono diversi strumenti per farlo, i modi più comuni sono [Albano 13]:

- *Rapporto tradizionale*, che organizza il risultato con colonne di dati, intestazioni e uno o più livelli di dati di riepilogo parziali;
- *Tabella a doppia entrata*, che mostrano le aggregazioni delle misure rispetto ai valori delle dimensioni lungo gli assi cartesiani. Aggiungendo e togliendo dimensioni si ottengono rispettivamente le operazioni di *drill-down* e *roll-up*.
- *Grafici* di natura diversa, come istogrammi, diagrammi lineari, diagrammi a torta ecc.

Inoltre, si identificano tre modalità di attuazione del supporto alle decisioni:

- Reporting: il livello più basso di supporto alle decisioni, rappresenta il punto di partenza per la rappresentazione di informazioni significative per i manager.
- Analisi Multidimensionale: analisi interattiva delle informazioni raccolte nel data warehouse per mezzo di strumenti, come Excel, che lascino un certo grado di libertà all'utente finale.
- Analisi esplorativa: rappresenta una fase in cui le tecniche esplorative sono utilizzate al fine di estrarre modelli utilizzando algoritmi di data mining.

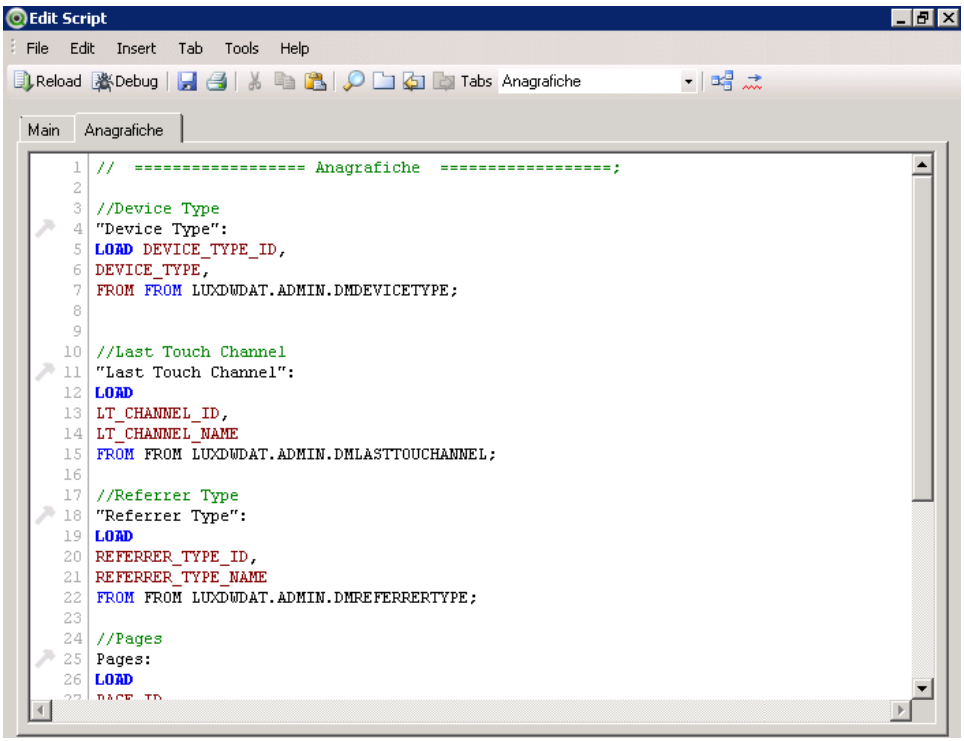


Figura 8.1: Esempio di caricamento delle anagrafiche

8.2 Caricamento dei dati

Al termine dell'esecuzione dei Job di Data Services vengono eseguiti gli script QlikView di caricamento dei dati. Ognuno di essi identifica la sorgente dei dati, le tabelle e i campi che devono essere caricati nella memoria virtuale. In particolare, nell'ambito di questo progetto, abbiamo:

- script di caricamento tabelle delle anagrafiche;
- script di caricamento tabella del fatto degli ordini di vendita;
- script di caricamento tabella del fatto clickstream.

Sono disponibili, in Figura 8.1 e 8.2, degli estratti del codice utilizzato per il caricamento delle tabelle di anagrafica e della tabella del fatti degli ordini di vendita. I join tra le tabelle sono garantiti dalla corretta nomenclatura

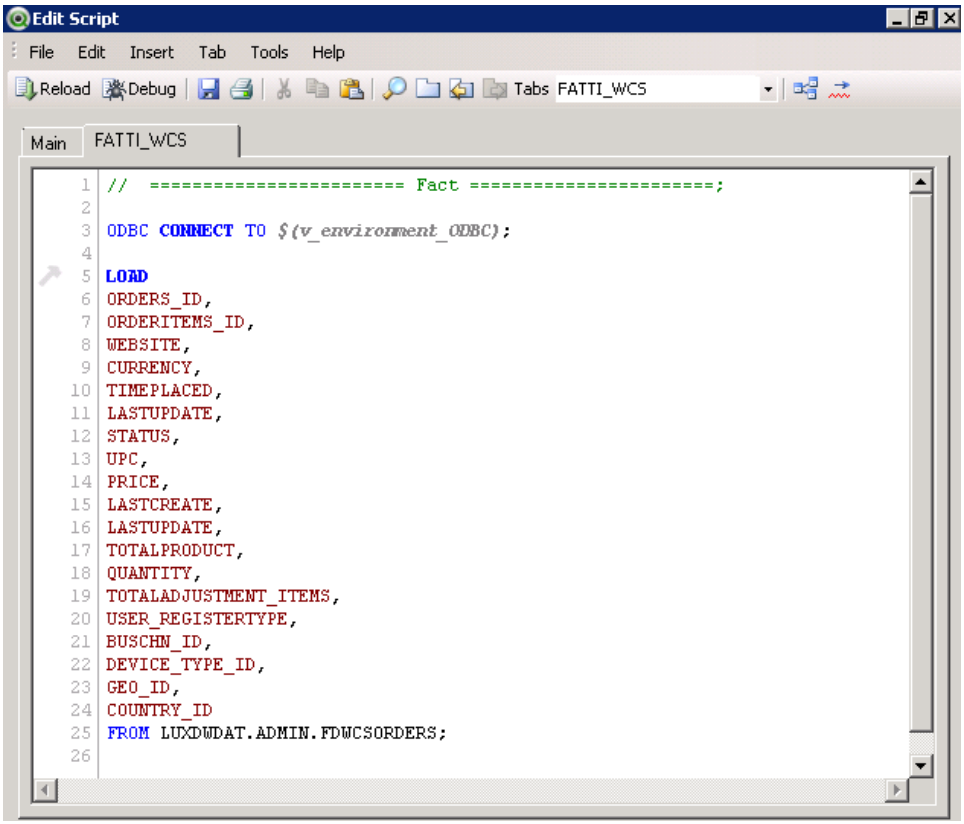


Figura 8.2: Esempio di caricamento della tabella dei fatti

dei campi. Come illustrato durante l'introduzione dello strumento (si veda Sezione 6.3), Qlikview adotta una logica associativa, che mette in relazione le tabelle sulla base dell'uguaglianza dei nomi dei campi. La mancanza di attenzione nell'assegnazione corretta dei nomi potrebbe causare join errati tra tabelle e spreco di memoria. Questo costituisce uno degli svantaggi principali di questa tecnica.

8.3 Funzionalità del sistema di reportistica

8.3.1 Esplorazione dei dati

Come riportato in [Garcia 12], il modello di dati gestito in memoria di QV, permette di analizzare i dati sia a livello aggregato che di massimo dettaglio. In aggiunta, le associazioni fra i dati vengono mappate automaticamente e rispondono alle selezioni dell'utente.

In QV l'analisi è eseguita su fogli navigabili tramite etichette. Ogni foglio può contenere più oggetti (caselle di riepilogo, grafici, tabelle ecc.) per analizzare il modello dei dati sottostanti. Tutti i fogli sono collegati tra loro per cui le selezioni effettuate su un foglio hanno effetto sui restanti.

In QV i risultati di una selezione possono essere visualizzati in un grafico. Di norma un grafico contiene una o più espressioni che vengono ricalcolate ogni volta che si esegue una selezione. I risultati sono visualizzati come grafici a barre, a linee, in scala di colore, a dispersione ecc. I grafici sono interattivi, quindi è possibile effettuare selezioni o query direttamente puntando il mouse e facendo clic selezionando l'area di interesse. Così come accade con la rappresentazione grafica dei dati, i risultati di un'analisi possono essere visualizzati in una tabella, anch'essa interattiva.

Figura 8.3: Intestazione dello strumento

Di seguito è riportata la struttura di navigazione delle schede realizzate. L'elenco puntato corrisponde alla numerazione riportata in Figura 8.3.

1. Filtri applicabili sulle dimensioni.
2. Collegamenti alle altre schede.
3. Nome della scheda.
4. Il collegamento *Report Info* apre un pop-up con le informazioni sulla scheda correntemente visualizzato.
5. Il collegamento *Dizionario KPI* accede ad una scheda con informazioni sul KPI e le dimensioni (si veda Sezione successiva).

6. Il collegamento *About* apre un breve tutorial in formato video.

7. Informazioni sulla sessione corrente, contiene il nome dell'utente connesso e l'ultimo aggiornamento dei dati che si stanno visualizzando.

8. Visualizzazione dei filtri attivi.

Quando l'utente clicca sul nome di una dimensione si apre un elenco dei valori ad essa associati e selezionabili. I valori della lista sono evidenziati con colori diversi: i verdi sono i valori selezionati, i bianchi sono i valori selezionabili e quelli in grigio sono i valori non selezionabili, in base ai filtri già attivi.

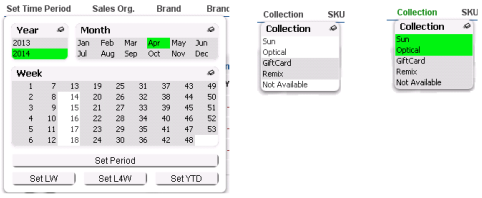


Figura 8.4: Selezione di una dimensione

L'utente può impostare un periodo predefinito o un periodo personalizzato (si veda Figura 8.4). I periodi predefiniti sono:

- LW ultima settimana;
- L4W ultime quattro settimane;
- YTD dal primo gennaio dell'anno corrente.

8.3.2 KPI Dictionary

In Figura 8.5 è rappresentata il *Dizionario dei KPI*, in questa scheda sono riportate tutte le informazioni disponibili sui KPI utilizzati dallo strumento. Il dizionario è raggiungibile da un collegamento inserito in ogni scheda.

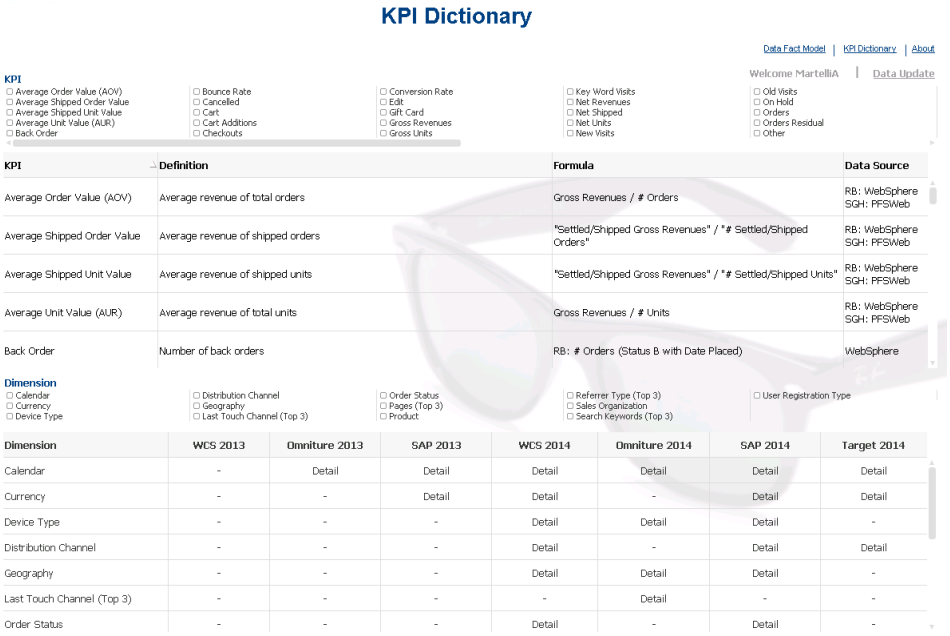


Figura 8.5: KPI Dictionary

8.4 Descrizione della reportistica

La reportistica è stata suddivisa come segue:

- **Home.** Prospettiva globale dei principali KPI;
- **Analisi degli ordini.** Sono presenti due schede, la prima consente di visualizzare gli ordini suddivisi per stati. Da qui l'utente può filtrare e ridurre l'insieme di ordini da analizzare, per poi approfondire sulla seconda scheda, contenente il dettaglio degli ordini;
- **Analisi del Prodotto.** L'analisi del prodotto presenta i KPI dei prodotti, classificandoli in base ad un insieme di misure definito dall'utente;
- **Analisi degli Utenti.** L'ultima scheda fornisce una visione grafica degli indicatori chiave di performance inerente al comportamento degli utenti all'interno del sito di commercio elettronico.

8.5 Home

La scheda principale mostra i valori dei principali KPI.

- Informazioni sulle visite, visitatori unici e conversion rate;
- informazioni sul numero di ordini e del valore medio di quelli spediti in termini di prezzo e di quantità;
- informazioni sui ricavi netti, lordi e sulle unità nette vendute e lorde (resi compresi).

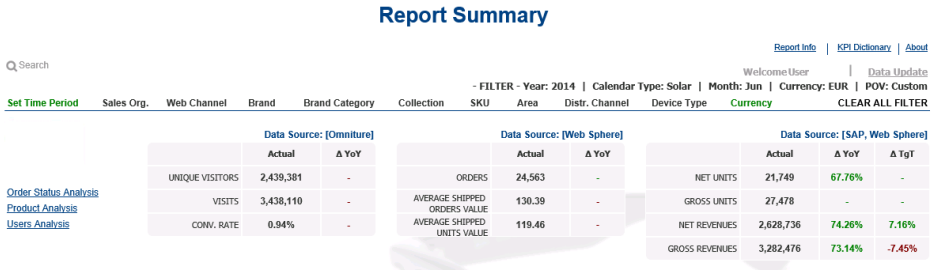


Figura 8.6: Report Principale

8.6 Analisi sugli stati dell'ordine

Diagramma di flusso degli ordini In ogni riquadro è inserito uno stato, il numero dei relativi ordini e la percentuale rispetto al totale di quest'ultimi. Alla selezione di un riquadro specifico, l'intero report viene filtrato in base allo stato selezionato. La tabella in basso a sinistra in Figura 8.7, contiene tutte le informazioni relative agli ordini appartenenti ad altri stati meno utili ai fini dell'analisi. La tabella subito a destra, presenta il totale degli ordini ed il valore medio di essi in base al prezzo e alle unità ordinate.

Order status variance Il grafico a barre orizzontali presenta la variazione degli ordini segmentati per stato rispetto al periodo di tempo precedente. L'utente può commutare tra l'analisi per anno, per mese o settimana.

Total order by week La curva rappresentata nel grafico mostra l'andamento del numero totale di ordini per settimana.

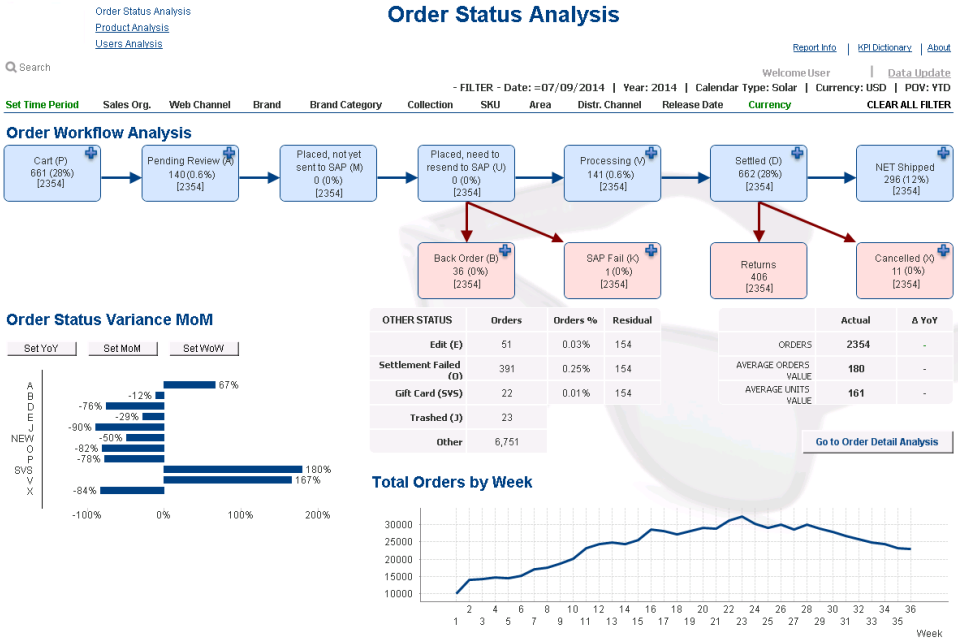


Figura 8.7: Analisi sugli stati degli ordini

8.7 Analisi dettagliata degli ordini

La scheda contiene le informazioni degli ordini nel massimo dettaglio (si veda Figura 8.8). Essa è dotata di un riquadro dedicato all'analisi libera per visualizzare in tabella solo le informazioni desiderate. Per aggiungere o rimuovere una dimensione o una misura, l'utente deve selezionare o deselezionare la casella relativa. L'analisi dettagliata sugli ordini ha lo scopo di fornire un report tabellare operativo. L'utente, infatti, possiede un ampio grado di libertà nella scelta delle dimensioni in considerazione di un numero di analisi potenzialmente maggiore rispetto a quelle espresse dalla specifica dei requisiti (si veda Capitolo 3).

8.8 Analisi del prodotto

La scheda contiene le informazioni sui prodotti nel massimo dettaglio. Al suo interno è presente un filtro sulla *release date* (vedi Figura 8.9), ovvero sulla data di lancio del prodotto. In questo modo, l'utente può definire un insieme di prodotti messi in vendita nello stesso periodo, allo scopo di effettuare un'analisi più mirata a valutare l'impatto del prodotto in termini di conversione (conversion rate).

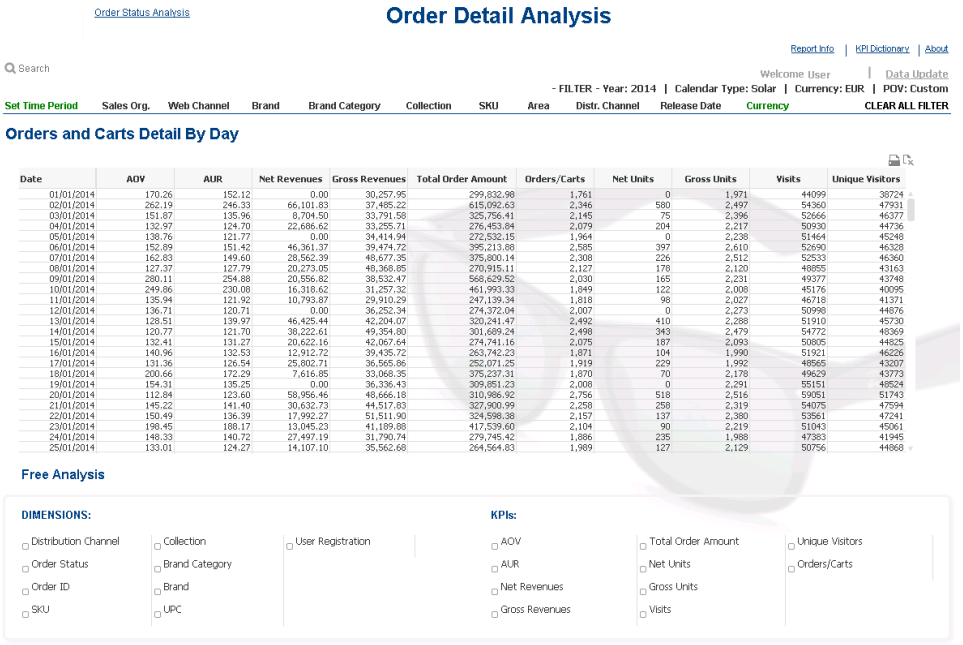


Figura 8.8: Analisi dettagliata degli ordini

Si descrivono nel seguito le sezioni da cui è composta la scheda sull'analisi del prodotto.

Product Conversion Funnel Il *Product Conversion Funnel* agevola l'analisi sui fattori che concorrono al calcolo del conversion rate e le principali operazioni effettuate dall'utente che lo portano alla generazione di un acquisto. La sua definizione ad *imbuto* fa capire in che modo ci sarà una decimazione al suo interno tra i visitatori e le vendite.

Grafico a bolle Il grafico a bolle è una variante di un grafico a dispersione in cui i punti dati vengono sostituiti con bolle e la loro ampiezza rappresenta una dimensione aggiuntiva. La posizione dell'elemento è determinata, in ordinata, dal numero totale di visitatori unici e, in ascissa, dall'importo totale delle entrate nette. La dimensione delle *bolle* rappresenta il totale degli ordini di quel prodotto. Alla selezione di una bolla, sia il Product Conversion Funell sia la tabella sottostante vengono filtrate per agevolare un'analisi mirata.

Free Analysis Il report tabellare presenta una visione dettagliata di tutte le misure e dimensioni disponibili associate al singolo prodotto. La sezione dedicata all'analisi libera, permette all'utente di aggiungere o rimuovere di-

mensioni a seconda delle sue esigenze. Anche in questo caso, il numero di analisi è potenzialmente maggiore rispetto a quelle espresse dalla specifica dei requisiti.

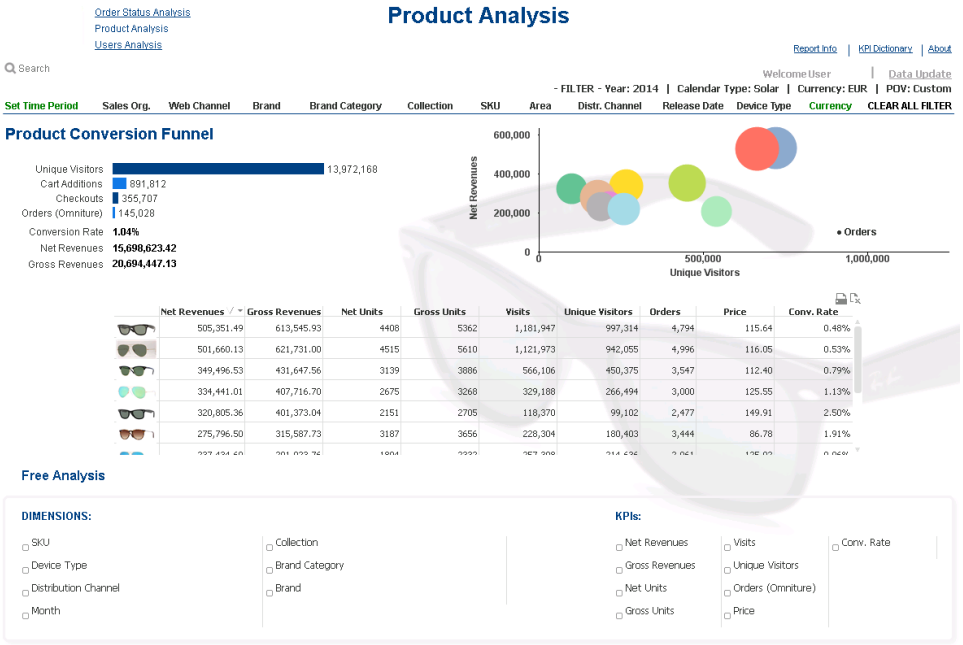


Figura 8.9: Analisi sul Prodotto

8.9 Analisi dell'utente

Nella scheda dedicata all'utente (si veda Figura 8.10), sono riportate tutte le informazioni relative al suo comportamento all'interno del sito di commercio elettronico. Di seguito vengono descritte le varie sezioni di cui è composta la scheda.

Overview. La prima sezione fornisce una *overview* su alcune metriche riguardanti:

- visite;
- visitatori unici;
- tasso di conversione;
- visitatori che hanno effettuato la loro prima visita;
- vecchi visitatori;

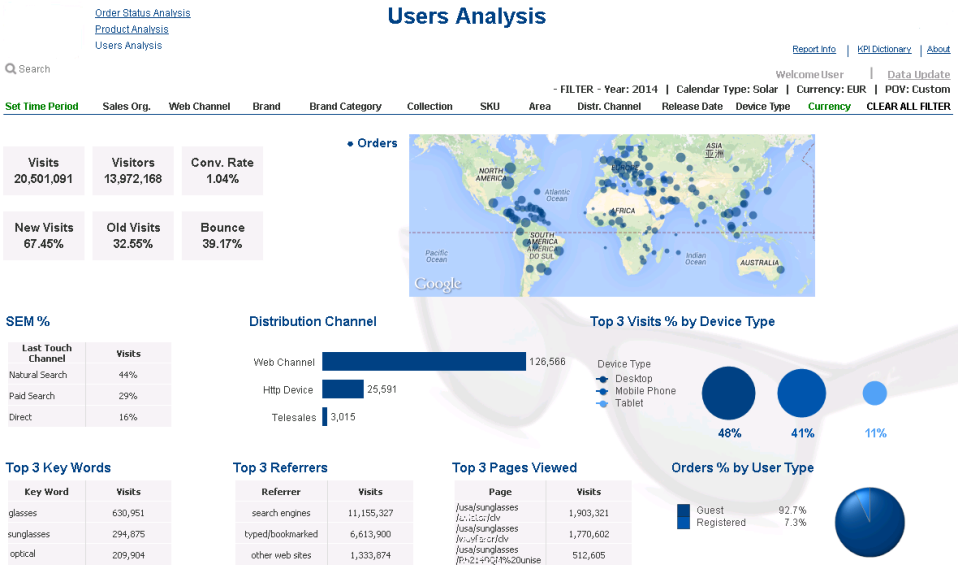


Figura 8.10: Analisi sull'Utente

- visite di rimbalzo.

Mappa Geografica. Sezione dedicata alla rappresentazione geografica del numero di ordini generati nel mondo. L'utente può selezionare il punto interessato e filtrare l'intera scheda.

% Search Engines Marketing. La tabella mostra la percentuale delle visite suddivisa per fonte di traffico:

- **Direct:** visitatori provenienti da link non riconducibili a nessun dominio;
- **Paid:** visitatori provenienti da motori di ricerca per risultati sponsorizzati;
- **Natural Search:** visitatori provenienti da motori di ricerca per risultati non sponsorizzati.

Il report ha lo scopo di suggerire azioni da intraprendere nel futuro. Se il sito web è ben ottimizzato il valore *Natural search* deve mantenere valori elevati.

Distribution channel. Il grafico a barre rappresenta la segmentazione dei visitatori in base al canale di distribuzione utilizzato per generare un ordine:

- **Web channel.** Ordini generati attraverso un browser web (laptop, PC, game console ecc.);
- **Telesales.** Ordini generati tramite televendita (ordini presenti unicamente in Web Sphere);
- **HTTP Device.** Ordini generati tramite smartphone, tabphone o tablet.

Top 3 Visits % by Device Type. Il grafico mostra una segmentazione del numero dei visitatori in base al tipo di dispositivo utilizzato per accedere al sito Web. L'utente può selezionare una delle tre rappresentazioni per filtrare l'intero report in base alla selezione scelta. L'analisi dei visitatori sui vari dispositivi, permette di valutare la facilità di utilizzo del sito accedendovi, per esempio, tramite un dispositivo mobile.

Top 3 Search Keywords. La tabella mostra le parole chiave con cui i visitatori, tramite un motore di ricerca, sono arrivati nel sito. Lo scopo di questo report è di fornire informazioni sulle parole chiave che portano visite di valore al sito, ad esempio confrontando, per ogni chiave di ricerca, il tasso di rimbalzo o di conversione.

Top 3 Referrers. Il report mostra la tipologia di link che hanno generato una visita. Tra le tipologie di refferrer abbiamo:

- **Search engines:** link generati dai risultati di un motore di ricerca;
- **Typed/Bookmarked:** link salvati nei preferiti del browser;
- **Other web sites:** link richiamati da altri siti web.

Top 3 Page. Mostra le pagine più popolari del sito. Le attività finalizzate ad ottenere una migliore rilevazione, analisi e lettura del sito web da parte dei motori di ricerca deve garantire che le pagine mostrate in tabella siano delle *landing page*, ovvero pagine web specificamente strutturate che il visitatore raggiunge dopo aver cliccato un link o una pubblicità.

Orders % by User Type. Il grafico a torta mostra la segmentazione degli utenti suddividendoli per registrati ed utenti ospiti. Il grafico ha l'obiettivo, di offrire un confronto fra il diverso comportamento delle due classi di utenti rispetto alle metriche visualizzate all'interno della scheda.

Conclusioni

In questo lavoro di tesi sono state approfondite le tematiche legate alla Web Analytics e al Data Warehousing. L'obiettivo principale è stato disegnare e realizzare un Data Warehouse a supporto di processi aziendali di controllo delle vendite e analisi del cliente nel commercio elettronico.

Lo sviluppo della funzione del marketing nelle imprese è parte fondamentale di una strategia di business che viene definita *proattiva*. In quest'ottica l'impresa ha un ruolo propositivo nei confronti dei bisogni del mercato.

Attraverso l'integrazione della Web Analytics nella strategia di marketing online, l'impresa ha la possibilià di valutare, in tempo reale, la sua efficacia migliorando l'esperienza online dell'utente. Tale strategia si pone l'obiettivo di incentivare il traffico sul sito, influenzare il comportamento dei visitatori ed accrescere le potenzialità di un loro acquisto.

Di conseguenza, le imprese che competono in un mercato mutevole necessitano di uno strumento che raccolga i dati provenienti da sorgenti eterogenee e li trasformi per produrre un'informazione volta al miglioramento della capacità decisionale.

A tale scopo nasce il Data Warehouse che costituisce l'oggetto di questa tesi. Ciò che ha richiesto maggior attenzione è stata l'attività di trasformazione dei dati, allo scopo di renderli aderenti alla logica di business del sistema di analisi. L'operazione di pulizia e consolidazione dei dati disponibili ha garantito la realizzazione di una base dati integrata di buona qualità.

In particolare, l'integrazione di informazioni eterogenee in una base di dati unica ha permesso di realizzare un'anagrafica prodotto completa e corretta. Questo ha permesso di sollevare l'utente di business dall'esecuzione di operazioni manuali di ricostruzione dell'anagrafica e di recuperare i dati collocati in ambienti differenti.

La metodologia di progettazione del Data Warehouse proposta ed applicata, riportata in [Albano 13], si è dimostrata all'altezza di affrontare il problema della gestione di grandi quantità di dati, in modo strutturato ed efficace.

L'approccio metodologico applicato non ha trascurato il contatto con il cliente al fine ultimo di soddisfare le sue esigenze nella maniera più semplice e

intelligente possibile, cercando di percepire le sue future potenziali necessità. A questo scopo è risultato essere indispensabile possedere una padronanza del dominio di business dell'utente, poiché solamente in questo modo si è in grado di comprendere i suoi reali bisogni e si può migliorare la qualità del suo lavoro. Le esigenze informative sono state soddisfatte ed i report realizzati sono stati ritenuti capaci di rispondere a tutti i requisiti di analisi rilevati in fase iniziale.

Il progetto ha rappresentato un punto di partenza per altre interessanti attività, attualmente in fase di sviluppo. Il nuovo obiettivo è quello di estendere le analisi del cliente ad altri sistemi di commercio elettronico appartenenti alla stessa azienda e di arricchire l'anagrafica prodotto includendo altre caratteristiche (colore asta, tipo di frontale, colore lenti, ecc.).

Il lavoro realizzato in ICONSULTING S.p.A. mi ha consentito di osservare l'applicazione concreta dei fondamenti teorici del processo di Data Warehousing per la soluzione dei problemi legati alla natura dei dati, alle modalità del loro trattamento e all'ambiente di sviluppo messo a disposizione dal cliente. Ció ha permesso di approfondire le mie conoscenze, fino ad allora esclusivamente di tipo accademico e di applicarle ad un caso reale.

ICONSULTING S.p.A. ha fornito un ambiente sereno per la realizzazione del progetto e una fiducia appagante che ha reso l'esperienza indimenticabile.

Bibliografia

[1] Casaleggio Associati, *Strategie di Rete, Rapporto Aprile 2014, L'ecommerce in italia*, ⟨http://www.casaleggio.it/e-commerce/⟩, 2014

[2] Kaushik A., *Web Analytics Definitions*, Web Analytics Association, 2007

[3] Kaushik A., *Web Analytics 2.0*, John Wiley & Sons, 2014

[4] Kimball R., *The Data Warehouse Toolkit, Pratical techniques for building dimensional data warehouse*, John Wiley & Sons, 1996

[5] Kimball R., *Data Webhouse Toolkit*, John Wiley & Sons, 2000

[6] Albano A., *Decision Support Databases Essentials*, Universitá di Pisa, 2013

[7] Iconsulting, *I sistemi a supporto delle decisioni. La teoria e la metodologia di riferimento*, http://wiki.iconsulting.biz, 2014

[8] Sap, *SAP Business Objects Data Services Tutorial*, ⟨http://help.sap.com/businessobject⟩, 2014

[9] Di Miguel Garcia, Barry Harmsen, *QlikView 11 for Developers*, Packt Publishing Ltd, 2012

[10] Netezza Corporation, *L'architettura dell'appliance Netezza*, ⟨http://www-01.ibm.com/software/⟩, 2009